TERMINOLOGY EXTRACTION AND ALIGNMENT FOR THE TRANSLATION INDUSTRY

Andraž Repar

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Asst. Prof. Senja Pollak, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Marko Robnik Šikonja, Chair, Jožef Stefan International Postgraduate School and Faculty of Computer and Information Science of the University of Ljubljana, Ljubljana, Slovenia
Prof. Dr. Špela Vintar, Member, Faculty of Arts of the University of Ljubljana, Ljubljana, Slovenia
Prof. Dr. Antoine Doucet, Member, University of La Rochelle, La Rochelle, France, and Faculty of Computer and Information Science, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Andraž Repar

TERMINOLOGY EXTRACTION AND ALIGNMENT FOR THE TRANSLATION INDUSTRY

Doctoral Dissertation

LUŠČENJE IN PORAVNAVA TERMINOLOGIJE ZA PREVAJALSKO INDUSTRIJO

Doktorska disertacija

Supervisor: Asst. Prof. Senja Pollak

Ljubljana, Slovenia, January 2025

Acknowledgments

I would like to thank my colleagues from the Jožef Stefan Institute, in particular Matej Martinc, Vid Podpečan, and Tran Hanh, for their support and collaboration on various papers. I am also grateful to Nada Lavrač for encouraging me to start a PhD in the first place, and for all her help with writing papers, discussing ideas and guiding me throughout my research. Finally, my deepest thanks go to my supervisor, Senja Pollak, for her guidance and valuable feedback, without which this work would not have been possible.

The work was partially performed in the scope of the projects "Cross-lingual and cross-domain methods for Terminology Extraction and Alignment" (grant number BI-FR/23-24-PROTEUS006), bilateral project funded by the Slovenian Research Agency (ARIS) scheme PROTEUS, the ARIS funded project "TermFrame - Terminology and Knowledge Frames across Languages" (No. J6-9372), the project "Development of Slovene in Digital Environment" (RSDO) funded by the Ministry of Culture of the Republic of Slovenia, the project "EMBEDDIA: Cross-Lingual Embeddings for Less-Represented Languages in European News Media" (EC-funded H2020 RIA project, grant number No. 825153).

Abstract

This PhD dissertation focuses on improving terminology extraction and alignment for applications in the translation industry. It explores three key use cases where these techniques benefit language professionals: creating client-specific terminology lists from large parallel corpora (i.e. translation memories), building domain-specific terminology resources from comparable corpora, and identifying important domain-specific terms in source documents prior to translation.

The research starts with the task of bilingual terminology alignment from parallel corpora found in translation memories. The main contribution is a novel approach called Phrase-Table-Based Alignment, which uses phrase tables from statistical machine translation to align terms with greater accuracy at the sub-sentence level. Additionally, the dissertation introduces TermEnsembler, a terminology extraction and alignment system developed for an industry partner which utilizes an ensemble learning approach combining seven different alignment methods via an evolutionary algorithm to find the best combination for optimal results. TermEnsembler was tested on three industry-specific domains (Financial, Information Technology, and Automotive) using a precision-focused evaluation of the top-ranked outputs.

Next, the dissertation addresses bilingual terminology alignment from comparable corpora. We start by replicating an existing machine-learning approach and then incorporate additional dictionary-based (features that leverage bilingual dictionaries or word alignments) and cognate-based (features that leverage words with shared etymological origins that exhibit similarities across languages) features. Later work expanded on this by implementing word alignments based on cross-lingual word embeddings and sentence embeddings. The methods were evaluated on the Eurovoc thesaurus, a multilingual thesaurus of EU-related terminology, using standard evaluation metrics, as well as adapted for keyword alignment in the media industry.

Finally, the dissertation explores two approaches to monolingual terminology extraction from specialized corpora with a special focus on the Slovenian language. The first is a machine-learning method combining statistical, linguistic, and contextual features derived from contextual embeddings. It employs feature engineering to capture termhood and unithood characteristics, resulting in improved precision and recall metrics over traditional approaches. The second approach, which surpasses the performance of the machine-learning approach, utilizes transformer-based models for sequence labeling, assigning a label to each token in a text sequence. Both approaches were evaluated on the RSDO5 dataset, a specialized dataset with four domains created specifically for terminology extraction evaluation.

These findings provide practical improvements for the translation industry. By adopting these new approaches, the industry can effectively leverage the existing language resources at their disposal to achieve more accurate and consistent specialized terminology. Overall, they offer a step forward in making terminology extraction and alignment more reliable and efficient, supporting better outcomes for language professionals and their clients.

Povzetek

Ta doktorska disertacija obravnava luščenje in poravnavo terminologije v prevajalski industriji. Osredotoča se na tri ključne primere rabe, kjer te tehnike koristijo prevajalcem in prevajalskim podjetjem: generiranje glosarjev za posamezne naročnike iz velikih paralelnih korpusov (tj. prevajalskih baz), gradnjo terminoloških virov iz primerljivih korpusov in prepoznavanje relevantnih domenskih terminov v izvornih dokumentih pred začetkom prevajanja.

V prvi fazi se posvetimo dvojezični poravnavi terminov v prevajalskih bazah, zbirkah prevodov, ki so poravnani na nivoju stavkov. Glavni prispevek je pristop s pomočjo tabel besednih zvez, ki se uporabljajo pri statističnem strojnem prevajanju in s katerim dosežemo večjo natančnost pri poravnavi terminov in besednih zvez znotraj stavkov. Poleg tega disertacija vključuje opis sistema za luščenje in poravnavo terminologije TermEnsembler, ki je bil razvit za naročnika raziskave. Sistem s pomočjo ansambelskega učenja združi rezultate sedmih metod za poravnavo terminov in z evolucijskim algoritmom poišče najboljšo možno kombinacijo. Rezultati sistema TermEnsembler so bili evalvirani na treh domenah (finance, informacijska tehnologija in avtomobilska industrija) s poudarkom na natančnosti najvišje rangiranih kandidatov.

Nato se osredotočimo na dvojezično poravnavo terminov v primerljivih korpusih. Najprej repliciramo obstoječ pristop, ki za poravnavo uporablja strojno učenje, in nato modelu dodamo dva tipa dodatnih značilk: slovarske, ki za poravnavo besed uporabljajo dvojezične slovarje in glosarje, in take, ki temeljijo na besedah skupnega etimološkega izvora, zaradi česar zvenijo podobno v več jezikih. Pozneje dodamo še nove poravnave besed s pomočjo medjezikovnih vektorskih vložitev in stavčnih vložitev. Razviti pristopi so evalvirani s pomočjo večjezičnega tezavra EU-terminologije Eurovoc, polega tega pa so bili tudi prilagojeni za poravnavo ključnih besed v medijski industriji.

V zadnji fazi se osredotočimo na dve metodi luščenja terminologije iz specializiranih korpusov za slovenščino. Prva metoda temelji na klasičnem strojnem učenju in združuje statistične, lingvistične in kontekstne značilnosti. Z generiranimi značilnostmi zajamemo tipične lastnosti terminov in dosežemo boljšo natančnost ter priklic v primerjavi s tradicionalnimi pristopi. Druga metoda, ki doseže še boljše rezultate, uporablja modele na osnovi arhitekture transformer. Vsaki enoti v besedilu pripiše oznako in tako luščenje terminologije obravnava kot problem označevanja zaporedij. Obe metodi sta evalvirani s pomočjo korpusa RSDO5, ki vsebuje štiri domene in je bil ustvarjen posebej za evalvacijo luščenja terminologije.

Razviti pristopi ponujajo praktične izboljšave za prevajalsko industrijo. Z njihovo uvedbo lahko industrija učinkovito izkoristi obstoječe jezikovne vire, ki so ji na voljo, ter zagotovi natančnejšo in doslednejšo uporabo specializirane terminologije. Na ta način prispevajo k večji kakovosti prevodov ter izboljšujejo kakovost storitev, ki jih prevajalci in prevajalska podjetja nudijo svojim naročnikom.

Contents

List of Figures

List of Tables

Abbreviations

SVM ... support vector machine
UD ... universal dependencies
TMX ... Translation Memory eXchange
CAT ... computer-assisted translation
TM ... translation memory
LLM ... large language model
PBMT ... phrase-based machine translation
NMT ... neural machine translation
ATE ... automated terminology extraction
POS ... part of speech
CT ... candidate term
SWU ... single word unit
MWU ... multi word unit
GA ... genetic algorithm
RAG ... retrieval augmented generation
SOTA ... state-of-the-art

Chapter 1

Introduction

In the introductory chapter, we first describe the problem of terminology extraction and alignment addressed in the thesis (Section 1.1), discuss related work (Section 1.2) and propose the goals we aim to achieve (Section 1.3), with specific focus on the translation industry. We present the initial hypotheses on which the thesis is based in Section 1.4. Finally, we conclude by explaining the scientific contributions the thesis offers in Section 1.5 and by presenting the structural overview in Section 1.6.

1.1 Background and Problem Definition

This thesis addresses two main research problems: terminology extraction and terminology alignment, which are of crucial importance for semi-automated terminology management in contemporary translation industry, i.e. companies that provide translation and other related language services.

The term "translation industry" describes collectively the companies that provide translation and other related language services. This industry has been steadily growing since the start of the 1990s and the increase in the volume of translated words brought along the need to streamline the translation process with automated solutions. Unlike many processes that have been effectively automated, terminology remains one of the main problem areas in the translation industry. For example, a report by SDL, a market leader in translation and terminology management software solutions, showed that among 140 companies, 51 percent of the respondents did not have a terminology management process in place, while a survey by Meex and Straub (2016) showed that among 800 respondents, 89.5 percent often or constantly experience that different organizational areas or employees use different terms for the same concept and that 51.9 percent of employees often or constantly cannot understand terms immediately.

Professional translation takes place inside specialized editors, which are part of computer-assisted translation (CAT) software suites. A crucial pre-processing step is segmentation – a document, for example a PDF or Word file, is segmented into smaller segments. For the most part, these correspond to individual sentences, but they can also be smaller or larger units (depending on the quality of the segmentation algorithm). Translators then translate the document segment by segment storing each segment pair (i.e. source and target) into the translation memory[1] (TM). If a segment has already been translated before, the CAT tool will recall it from the TM, thus sparing the translator the time and effort needed to translate it from scratch. As a by-product of this process, translation companies have large parallel corpora at their disposal.

[1] A translation memory is a sentence-based parallel corpus compiled while translated using a computer-assisted translation tool.

Over the past decade, the translation industry has seen a major shift. From phrase-based machine translation (PBMT) to neural machine translation (NMT) and, more recently, large language models (LLM), the role of translators has changed dramatically. These new technologies help translators work faster and improve quality—if they know how to use them well. However, these technologies often fall short in one key area: terminology. No matter how linguistically polished or consistent a translation is, it doesn't work if the terminology is wrong, especially in critical fields like legal documents, medical device instructions, or patents. Relying too heavily on machine translation or LLMs for terminology can be risky[2]. These systems pull from vast datasets, much of it from the internet, which includes a surprising amount of machine-translated content of questionable quality (Thompson et al., 2024). In contrast, the best terminology often comes from small glossaries and translation memories that have been carefully built over time by skilled translators and specialized agencies.

A common example of a terminological challenge is the English term "heat exchanger", which can be translated into Slovenian in several ways:

- DeepL on September 17, 2024: *toplotni izmenjevalnik* (alternatives: *izmenjevalnik toplote*, *toplotni izmenjevalec*)
- Google Translate on September 17, 2024: *toplotni izmenjevalnik* (alternative: *izmenjevalec toplote*)
- chatGPT on September 17, 2024: *toplotni izmenjevalec*[3]

All of these translations are valid in certain contexts, and some can be used interchangeably, but a domain expert would note key differences (e.g., see https://lahde.fs.uni-lj.si/prenosnik-toplote-ali-toplotni-izmenjevalec/). *Toplotni izmenjevalnik* and its variants (*izmenjevalnik toplote*, *toplotni izmenjevalec*, *izmenjevalec toplote*) typically refer to a complete product, while *toplotni prenosnik* (or *prenosnik toplote*), which was suggested by chatGPT when asking for alternatives, tends to describe a component within a larger system. Translation clients may have their own preferences, which translators must consider. However, in the context of ventilation devices with heat recovery, the most commonly used Slovenian term is *rekuperator*, which none of the above systems suggested. When asked, "what about in the context of ventilation devices with heat recovery," chatGPT did offer the correct term, but this requires the translator to have some domain-specific knowledge.

Translation technology providers have not prioritized easing terminological work for translators, instead focusing on integrating machine translation into workflows and general workflow optimization. Aikwit[4], a boutique translation company from Slovenia, primarily uses Phrase[5] as its translation environment but also works with various other tools, including RWS Trados[6], MemoQ[7], and internal systems like Translation Workspace by Lionbridge[8] and GlobalLink by Transperfect[9], as well as software localization platforms

[2]It's important to note that the latest LLM-based translation systems have a clear advantage over older machine translation models. While traditional systems translate text sentence by sentence with little regard for the overall context, LLMs can consider a much larger portion of the text at once. This has the potential to produce translations that are more accurate, consistent, and contextually appropriate.

[3]The prompt was "translate heat exchanger into Slovenian." Asking for alternatives provided: *izmenjevalec toplote*, *toplotni prenosnik*, and *prenosnik toplote*.

[4]https://aikwit.com/. This company is given as an example, as the PhD candidate is one of the founders and knows in detail their work processes.

[5]https://phrase.com/

[6]https://www.trados.com/

[7]https://www.memoq.com/

[8]https://www.geoworkz.com/Product

[9]https://globallink.transperfect.com/

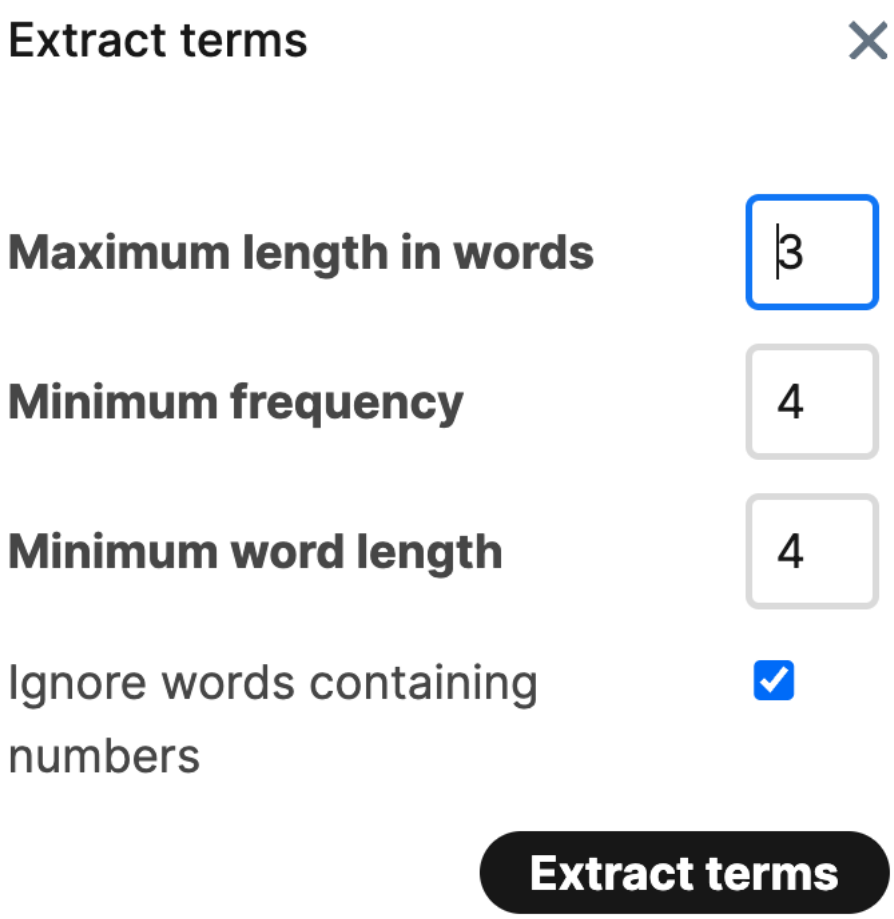

Figure 1.1: Terminology extraction functionality in the Phrase localization platform.

like Crowdin[10] and Lokalise[11]. While these tools offer glossary management and term extraction features, their functionality is limited and often produces low-quality results. For instance, Phrase's terminology extraction (see Figure 1.1) relies on simple frequency-based methods and Aikwit's experience suggests that other tools are similarly basic. Academic research into terminology has also overlooked the needs of translators, focusing on large datasets, while translators often work with small datasets or single documents.

In broad terms, this thesis addresses two interconnected problems. The first is **terminology extraction**, which refers to the *extraction of structured terminological knowledge from unstructured text in a single language*. The second is **terminology alignment**, defined as *the process of identifying translation pairs of terms across two (or more) languages*. These two tasks—terminology extraction and alignment—are often treated as steps within a single process, commonly described as bilingual or multilingual terminology extraction. Even in approaches labeled as "bilingual terminology extraction" or similar, the processes of terminology extraction and terminology alignment are generally treated separately.

[10] https://crowdin.com/
[11] https://lokalise.com/

1.1.1 Relevance of the problems in light of recent advancements

Given the rapid development of large language models (LLMs) in recent years, it is reasonable to question whether these problems remain relevant today. Instead of developing dedicated solutions for the challenges described above, LLMs may provide comparable or superior performance without the need for painstakingly crafted functionalities. As in many other fields, LLMs have revolutionized the translation industry; however, the issues of accurate and consistent terminology persist. In practical terms, beyond having access to accurate, vetted terms, translators often work in teams where consistent terminology usage is crucial. Ensuring that all team members can reference the same terminology database is vital for maintaining translation consistency across projects. Furthermore, terminology extraction tools play an important role in facilitating communication between translation companies and their clients. By extracting key terms and providing suggested translations, companies can present clients with comprehensive term lists for approval or clarification, thus streamlining the translation process. We believe that while LLMs will not completely resolve the challenges of terminology extraction and alignment, they will significantly enhance performance through methods such as prompting, retrieval-augmented generation, or other emerging techniques.

1.2 Related Work

In this section, we provide an overview of related work related to the two main research topics, i.e. terminology extraction and terminology alignment.

1.2.1 Terminology extraction

Terminology extraction refers to structuring terminological knowledge from unstructured text. According to ISO standard 087-1:2000 Terminology work — Vocabulary, terminology is a set of designations belonging to one special language. Automated terminology extraction (ATE) systems were traditionally classified as either statistical, linguistic or a combination of these two approaches. The linguistic approach utilizes the distinctive linguistic aspects of terms, most often their syntactic (i.e. part-of-speech) patterns. On the other hand, the statistical approach takes advantage of term frequencies in a corpus. However, most traditional systems are hybrid, using a combination of the two approaches. For example, Justeson and Katz (1995) first define part-of-speech (POS) patterns of terms and then use simple frequencies to filter the term candidates.

Many terminology extraction algorithms are based on the concepts of termhood and unithood defined by Kageura and Umino (1996): termhood is "the degree to which a stable lexical unit is related to some domain-specific concepts" and unithood is "the degree of strength or stability of syntagmatic combinations and collocations". Termhood-based statistical measures, such as Vintar (2010), function on a presumption that a term's relative frequency will be higher in domain-specific corpora than in the general language, while common statistical measures, such as mutual information (Daille et al., 1994), are used to measure unithood. These two approaches have been used as a basis of several hybrid systems, such as Termolator (Meyers et al., 2018) and TermEnsembler (Repar et al., 2019).

Another distinct approach is to utilize machine learning with feature engineering. It involves first extracting a set of term candidates, followed by the calculation of various features and training of a machine learning model, where term extraction is treated as a bilingual classification task. Various types of machine learning algorithms have been used, such as decision trees (Karan et al., 2012), rule induction (Foo & Merkel, 2010), k-nearest neighbours (Zadeh & Handschuh, 2014), support vector machines (Ljubešić et al., 2019)

and random forest (Rigouts Terryn et al., 2021a). The last two approaches are particularly relevant for our work.

The first approach, developed by Ljubešić et al. (2019) has long been considered as the state-of-the-art system for Slovene. It first extracts candidate terms (CT) with the CollTerm tool (Pinnis et al., 2012), which uses a complex language-specific set of term patterns originally developed for the SketchEngine terminology extraction module (Fišer et al., 2016). A total of 31 patterns were defined from unigrams up to four-grams and CTs (i.e. lemmatized versions of terms) with a frequency of 3 or more were considered. The resulting term lists were annotated by four annotators as either in-domain, out-of-domain, academic or irrelevant terms. The annotations were then used as training data for a machine learning approach with the following features: term frequency, 5 statistical measures from the SketchEngine terminology module (chi-square, dice, pointwise mutual information, t-score, tf-idf), C-value (Frantzi et al., 2000), candidate length, average token length, term pattern and context[12]. Additionally, oversampling was used to boost instances where the annotators were in agreement. Since some of the statistical measures can be calculated only for multi-word units, they trained separate classifiers for single-word (SWU) and multi-word (MWU) units. They evaluated the models in a cross-validation setting on the Slovene KAS dataset (Erjavec et al., 2018), obtaining F_1 scores of around 0.5 (i.e. when only the *irrelevant* annotation is considered negative and the remaining annotations are treated as valid terms).

The second approach, which also uses the machine learning paradigm with feature engineering was developed by Rigouts Terryn et al. (2021a). It first identifies the linguistic patterns of the annotated terms in the ACTER corpus (Rigouts Terryn, Hoste, & Lefever, 2020a) and then uses these patterns to identify term candidates. They generate 177 features in 6 subgroups: shape (e.g., number of tokens in a CT), linguistic (e.g., POS tag of the first token of the CT), frequency (e.g., relative frequency of the CT in a specialized corpus), statistical (various termhood/unithood measures), contextual (e.g., whether the CT occurs between parentheses or right before/after parentheses), variational (number of different variations of the CT). They experimented with several different classification algorithms in the *sklearn* Python library and obtained the best F_1 scores with the random forest classifier. In the setup with a held-out test set (3 domains are used for training, one for testing and the experiments are run 4 times, each time with a different test domain) they achieved F_1 scores between 0.338 and 0.436 for English, between 0.288 and 0.520 for French and between 0.361 and 0.616 for Dutch.

The development of word embeddings and deep learning models has significantly influenced the area of terminology extraction. One of the first attempts in the use of embeddings for terminology extraction was by Amjadian et al. (2016), who attempted to express unigram terms via a blend of local and global vectors. Various other strategies used non-contextual word embeddings. For example, Khan et al. (2016) leveraged word embeddings to calculate term similarity in a graph-based ranking method. Wang et al. (2016) created a co-training system with two neural networks to decide if a term is domain-specific or not. Zhang et al. (2017) used word embeddings to determine the semantic correlation of term candidates, aiding in the re-ranking of candidates produced by conventional term extraction techniques. Kucza et al. (2018a) identified term candidates via sequence labeling and both word-level and character-level embeddings. Gao and Yuan (2019) developed a nested term extraction classifier incorporating features from different word embedding models, both non-contextual and contextual. Contextual word representations such as

[12]Calculated by using a context-based SVM classifier with a linear kernel with features of the classifier being frequencies of tokens occurring in a 3-token window around all the occurrences of a term candidate in the respective document.

ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) can incorporate more term information, evidenced by the fact that the winning strategy in the TermEval2020 competition (Rigouts Terryn, Hoste, Drouin, et al., 2020) used the BERT model. TALN-LS2N (Hazem et al., 2020), the winning method for English and French, used BERT in a binary classification framework, where n-grams combined with a sentence served as an instance, and the classifier was required to decide if each n-gram within the sentence is a term or not. In contrast, the winning approach for Dutch mentioned in Rigouts Terryn, Hoste, Drouin, et al. (2020) used pretrained GloVe word embeddings as input for a bi-directional LSTM neural architecture.

Recently, a noteworthy shift in the effectiveness of terminology extraction techniques has been observed, driven by the adoption of a sequence-labeling approach using transformer-based large language models. This development follows the successful application of such approaches across a broad spectrum of other NLP tasks, including named entity recognition (Lample et al., 2016; H. T. H. Tran et al., 2021) and keyword extraction (Martinc et al., 2022). The study by Kucza et al. (2018b) was among the earliest to treat term extraction as a sequence-labeling task, an approach also explored in Conneau et al. (2020), Hazem et al. (2022), Rigouts Terryn et al. (2021b), Lang et al. (2021), and H. T. H. Tran, Martinc, Doucet, et al. (2022). The latter two methodologies have recorded F_1-scores nearing 0.6, representing a substantial enhancement over previous techniques. Many of these models also experiment with cross-lingual learning, obtaining promising results, and in some cases (e.g., H. T. H. Tran, Martinc, Doucet, et al. (2022)) even surpassing monolingual configurations. Finally, H. T. H. Tran, Martinc, Pelicon, et al. (2022) also shows that an ensemble of transformer approaches can achieve even better results. The papers by H. T. H. Tran, Martinc, Doucet, et al. (2022), H. T. H. Tran, Martinc, Pelicon, et al. (2022), as well as H. Tran et al. (2022), are also relevant for our work since they work with the Slovenian language.

Recent advancements in large language models (LLMs) have shown great potential for terminology extraction. Early work by Giguere (2023) is promising, indicating that LLMs can improve accuracy in certain aspects, while research by Banerjee et al. (2024) and H. T. H. Tran et al. (2024) show that in-context learning in a few-shot scenario can approach or even exceed the performance of fully supervised models, without the need for extensive data annotation and model training. In addition to in-context learning, generative LLMs can be fine-tuned for specific tasks, such as machine translation (MT), further enhancing their performance. However, more research is still needed to fully understand their capabilities for terminology extraction.

Finally, there have been several surveys, such as Kageura and Umino (2001), Conrado et al. (2014) (for Brazilian Portuguese) and most recently H. T. H. Tran et al. (2023) and shared tasks and workshops, such as Mustafa et al. (2006) and Rigouts Terryn, Hoste, Drouin, et al. (2020), focusing on terminology extraction.

1.2.2 Terminology alignment

Terminology alignment is the process of finding translation pairs of terms in two (or more) languages. Terminology extraction and alignment are usually considered two steps of one process (often described as bilingual or multilingual terminology extraction) and there are only a few papers that focus on terminology alignment exclusively, such as the paper by Aker et al. (2013). At the highest level, terminology alignment can be divided into alignment from comparable and alignment from parallel corpora, where parallel corpora are composed of source texts and their translations in one or more different languages aligned at the level of sentences, while comparable corpora are composed of monolingual texts collected from different languages using similar sampling techniques (McEnery et al.,

2006). For alignment of terms between the two languages, the methods typically utilize the idea that a term and its translation tend to occur in similar lexical contexts (Daille & Morin, 2005).

There are two distinct approaches to terminology extraction and alignment according to Foo (2012):

- Align-extract, where we first align single and multi-word units in and then extract the relevant terminology from a list of candidate term pairs, and
- Extract-align, where we first extract monolingual candidate terms from both sides of the corpus and then align the terms.

An advanced align-extract approach is proposed by Macken et al. (2013) utilizing a chunk-based alignment method to produce a list of candidate term pairs, which are then filtered using statistical methods. The extract-align approach is the more common of the two. Kupiec (1993) describes an algorithm for noun phrase extraction followed by alignment with a statistical estimation algorithm, achieving precision of 90 percent on the highest ranked candidate pairs. Vintar (2010) describes an extract-align approach named "bag-of-equivalents", where after monolingual extraction, the term pairs are aligned with the help of word alignment probabilities. Baisa et al. (2015) describe a frequency-based term alignment algorithm utilizing a variation of logDice to score the strength of the candidate term pair alignment. Haque et al. (2014) first generate monolingual candidate terms and then build a phrase table using the Moses toolkit (Koehn et al., 2007) and compare the extracted terms with the phrases in the table. Precision among the top 100 candidate term pairs often exceeds 90 percent. Aker et al. (2013) treat bilingual term alignment as a binary classification task, achieving good results. Hazem and Morin (2017) experimented with word embeddings used to augment bilingual terminology extraction from specialized comparable corpora.

Despite the problem of terminology alignment lending itself well to the binary classification task, there have been relatively few approaches utilizing machine learning. For example, similar to Aker et al. (2013), Baldwin and Tanaka (2004) generated corpus-based, dictionary-based and translation-based features and trained an SVM classifier to rank the translation candidates. Note that they only focus on multi-word noun phrases (noun + noun). A similar approach, again focusing on noun phrases, is also described by Cao and Li (2002). Nassirudin and Purwarianti (2015) reimplement the approach by Aker et al. (2013) for the Indonesian-Japanese language pair and further expand it with additional statistical features. In the best scenario, their accuracy, precision and recall all exceed 90% but the results are not directly comparable since Nassirudin and Purwarianti (2015) use 10-fold cross-validation, while Aker et al. (2013) use a held-out test set. In addition, Nassirudin and Purwarianti (2015) have a balanced test set while Aker et al. (2013) use a very unbalanced one (ratio of positive vs. negative examples 1:2000).

With the advent of word embeddings and deep learning models, new approaches to terminology alignment have also been developed. Conneau et al. (2018) developed a methodology for building bilingual dictionaries by aligning monolingual word embedding spaces using adversarial training. Artetxe et al. (2019) expanded on the same idea by using cross-lingual embeddings to build a phrase table, combine it with a language model, and use the resulting machine translation system to generate a synthetic parallel corpus, from they extracted the bilingual lexicon using statistical word alignment techniques. Shi et al. (2021) were able to get superior results by combining unsupervised bitext mining and unsupervised word alignment. Adjali, Morin, and Zweigenbaum (2022) exploit parallel corpora to build specialised comparable corpora which are then used to assess multi-word term alignment. Požár et al. (2022) propose 3 methods for term alignment: one based

on non-contextual term embeddings based on Bojanowski et al. (2016), another one based on unsupervised phrase-based machine translation and a third one based pretrained multilingual language models. Setha and Aliane (2023) propose a bilingual term alignment methodology for the Arabic-French language pair using contextual information based on the ELMo (Peters et al., 2018) embeddings.

1.3 Purpose and Goals of the Dissertation

As described in Section 1.1, having access to accurate and reliable terminology extraction and alignment algorithms would be of great benefit to language professionals working in the translation industry. There are three distinct use cases for terminology extraction and alignment

1. U1: terminology extraction and alignment of terms found in translation memories (i.e. parallel corpora) to build a list of client-specific terminology

2. U2: terminology extraction and alignment of terms found in comparable corpora to create domain-specific terminology

3. U3: terminology extraction from source documents before the start of the translation process to identify relevant terms in a domain-specific document (or a collection of documents)

The purpose of the dissertation is to develop algorithms for effective terminology extraction and alignment in the translation industry for the three use cases described above. We will attempt to combine traditional linguistic and statistical approaches with advanced machine learning and neural network-based natural language processing techniques to improve the performance of algorithms for monolingual terminology extraction, to address use cases U1, U2 and U3, algorithms for alignment of terms extracted from bilingual parallel corpora, to address use case U1, and algorithms for alignment of terms extracted from bilingual comparable corpora, to address use case U2.

In the experiments, we will include Slovenian language, on which state-of-the art approaches are rarely evaluated, by which we will contribute to the development of language technology solutions for Slovenian.

The goals of this dissertation are related to improving the terminology extraction and alignment process with a specific focus on the translation industry. The specific goals of the dissertation are as follows:

- G1: In relation to use case U1, we attempt to improve the performance of existing terminology alignment algorithms on parallel corpora from the translation industry. To do so, we develop a novel terminology alignment approach utilizing phrase tables and combine different approaches using evolutionary algorithms.

- G2: In relation to use case U2, we attempt to improve the performance of existing terminology alignment algorithms on comparable corpora. To do so, we replicate an existing supervised classification approach and adapt the approach to achieve satisfactory performance. In addition, we test new features based on cross-lingual embeddings.

- G3: In relation to use case U3, and as a step in U1 and U2, we attempt to improve the performance of existing terminology extraction algorithms from specialized corpora. To do so, we develop a novel terminology extraction approach utilizing contextual

word embeddings and integrate it with existing linguistic and statistical methods, as well as a second novel terminology extraction approach utilizing transformer models and sequence labeling.

1.4 Hypotheses

In this thesis, we will test the following hypotheses:

- We propose improving terminology alignment from parallel corpora through sub-sentence analysis and ensemble methods.
  - H1.1: While simple co-occurrence scores can provide good results in terminology alignment from parallel corpora, we hypothesize that terminology alignment performance can be significantly improved if we focus on the sub-sentence level via the construction of phrase tables and application of co-occurrence heuristics on the aligned phrases.
  - H1.2: While there is a large number of statistical approaches for term-alignment, we hypothesize that improved term alignment can be achieved by an ensemble approach, and that the complex problem of finding the optimal combination of different metrics can be effectively solved using evolutionary algorithms.
- We propose improving terminology alignment in comparable corpora with a binary classification approach.
  - H2: With two clearly defined classes (i.e. is a valid translation pair or not), terminology alignment between languages can be addressed as a classification task. We hypothesize that improved performance can be achieved by enhancing traditional terminology alignment features based on word alignment and cognate scores with novel features constructed from cross-lingual word embeddings.
- We propose leveraging contextual embeddings to enhance terminology extraction.
  - H3.1: Contextual word embedding models, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), have improved state-of-the-art performance of many NLP tasks. While some traditional terminology extraction approaches involve contextual information via co-occurrence statistics, we hypothesize that H3.1 applying contextual information from deep neural network models by exploiting the contextual differences in the domain and general corpora can help to improve terminology extraction performance.
  - H3.2: Terminology extraction using contextual embeddings in a sequence labeling setting can improve over traditional supervised classification approaches. We hypothesize that by using richer contextual information, these embeddings can better capture subtle meanings and domain-specific details leading to more accurate detection of terms in text.

1.5 Scientific Contributions

By attempting to improve terminology extraction and terminology alignment processes in the translation industry based on the four hypotheses from the previous section, we propose several novel efficient methods combining linguistic insights with natural language processing (NLP) and machine learning (ML) techniques. With reference to the three

distinct use cases and the respective goals of the thesis, the scientific contributions of this thesis are the following:

- A novel approach to bilingual term alignment from parallel corpora based on statistical machine translation phrase tables (Neubig et al., 2011), as well as a novel methodology using an evolutionary algorithm to combine solutions of an ensemble of elementary term alignment algorithms. This novel methodology is described in detail in Chapter 2 and was published in the following journal publication:

  Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., & Pollak, S. (2019). TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment. Terminology, 25(1), 93-120.

- A novel approach to bilingual term alignment from comparable corpora using cognates (i.e. words that look similar across languages), cross-lingual word embeddings and sentence embeddings to generate features in a machine learning model. We initially replicated an existing machine learning approach to terminology alignment (Aker et al., 2013) with a focus on reproducibility, but we later expanded the model with new features and tested it also for the purpuses of aligning keywords for media industry. The proposed methodology is covered in detail in Chapter 3 and was published in the following publications:

  Repar, A., Martinc, M., & Pollak, S. (2020). Reproduction, replication, analysis and adaptation of a term alignment approach. Language Resources and Evaluation, 54(3), 767-800.

  Repar, A., Martinc, M., Ulčar, M., & Pollak, S. (2021). Word-embedding based bilingual terminology alignment. Proceedings of Electronic lexicography in the 21st century: Post-editing lexicography, 408-417.

  Repar, A., Pollak, S., Ulčar, M., & Koloski, B. (2022). Fusion of linguistic, neural and sentence-transformer features for improved term alignment. Proceedings of the BUCC Workshop within LREC 2022, 61-66.

  Repar, A., & Shumakov, A. (2021). Aligning Estonian and Russian news industry keywords with the help of subtitle translations and an environmental thesaurus. Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, 71–75.

- A novel approach to monolingual terminology extraction involving a combination of statistical, linguistic and contextual features in a machine learning setting. Building on the machine learning approach that we adopted in our experiments with terminology alignment, we transform traditional frequency and POS-pattern-based aspects of terminology extraction into features of a binary classification model, and add novel features based on contextual word embeddings. The proposed methodology is covered in Chapter 4. The approach was presented at the 18th TOTh International Conference in 2024 and will be published in the conference proceedings.

- A novel approach to monolingual terminology extraction involving transformer models and sequence labeling. This approach is described in Chapter 4 and was published in the following publications[13]:

  Tran, H., Martinc, M., Repar, A., Doucet, A., & Pollak, S. (2022). A Transformer-based sequence-labeling approach to the Slovenian cross-domain automatic term ex-

[13]My main contributions to the paper were in evaluation, design and error analysis.

traction. Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia, 196-204.

Tran, H., Martinc, M., Repar, A., Ljubešić, N., Doucet, A., & Pollak, S. (2024). Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling?. Machine Learning, 113, 1-30.

1.6 Organization of the Thesis

This thesis is structured as follows.

In Chapter 2, we cover our approach to bilingual term alignment from parallel corpora, which was our initial focus. We start by introducing the topic in Section 2.1 and then describe our methodology based on phrase table alignment and evolutionary algorithms in Section 2.2 and results in Section 2.3, followed by the paper Repar et al. (2019).

In Chapter 3, we cover our work on bilingual term alignment from comparable corpora. After introducing the topic in Section 3.1, we describe several variants of our methodology based on a machine learning approach using statistical, linguistic and contextual features in Sections 3.2, 3.3, 3.4 and 3.5, followed by papers Repar et al. (2020), Repar et al. (2021), Repar et al. (2022) and Repar and Shumakov (2021).

In Chapter 4, we present our work in the area of terminology extraction from monolingual corpora. We first describe the dataset used in our experiments in Section 4.1, a machine learning approach to monolingual terminology extraction in Section 4.2 and a sequence-labeling approach to monolingual terminology extraction in Section 4.3, followed by the paper H. Tran et al. (2024).

We finish the thesis in Chapter 5 where we summarize our work and scientific contributions in Section 5.1. Section 5.2 contains a discussion on the strengths and weaknesses of the developed approaches and methodologies and Section 5.3 presents plans and ideas for future work.

Chapter 2

Bilingual Terminology Alignment From Parallel Corpora

This chapter presents a comprehensive solution for a terminology extraction and bilingual alignment from parallel corpora. It covers a specific use case in the translation industry where companies have accumulated large amounts of bilingual data over time due to the use of CAT tools in their workflows.

In Section 2.1, we introduce the topic and define the problem and in Section 2.2, we describe our approach, including the novel elements, to a terminology extraction and alignment workflow for the translation industry. The results are presented in Section 2.3. The relevant paper for this chapter is:

Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., & Pollak, S. (2019). TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment. Terminology, 25(1), 93-120.

The code for this part of the research is not publicly available, as it was developed exclusively for an industry client.

2.1 Introduction

As described in Chapter 1, translation companies have large parallel corpora (called "translation memories") populated over the years with hundreds or, in some cases, millions of segment pairs. CAT tools, i.e. specialized translation software used by professional translators, first segment each source document into sentences and these are then translated one after the other, resulting in aligned pairs of source and target segments. Highly organized companies have tens of hundreds of these translation memories organized by domain and/or client and these were the starting point of our research. Specifically, we wanted to investigate how to get from a collection of sentence pairs (i.e. translation memory) to a bilingual glossary of domain-specific terminology. For the initial extraction of terms we adapted an existing hybrid (i.e. statistical and linguistic) approach for terminology extraction and for alignment we developed a novel methodology, which was the main scientific contribution. The work was not limited to just research activities - we developed an actual working terminology management system with a module for semi-automated bilingual terminology extraction and alignment, where "semi-automatic" refers to the ability of the users to edit and correct the results of the automatic extraction and alignment processes.

2.2 Description of the Approach

To address the needs of the translation industry we propose a system for semi-automated terminology extraction and alignment, currently focusing on English and Slovenian. The system, developed for one of the largest language service providers in Southeast Europe, consists of:

- A concept-oriented terminology database, where all the data is stored, allowing import from and export into industry-standard terminology management formats.
- A terminology extraction workflow, including automated extraction or import of manually defined monolingual terminology, followed by a novel approach to term alignment utilizing an evolutionary algorithm to combine the results of several individual bilingual term alignment methods.
- A web interface for managing the database and controlling the extraction and alignment algorithms.
- Additional functionalities for extraction of good example sentences and identification of the domain in which the term is used.

We propose a novel Phrase-Table-Based Alignment (PTBA) method based on Pialign (Neubig et al., 2011), as well as a methodology using an evolutionary algorithm based on the DEAP framework (De Rainville et al., 2012) to combine the results of an ensemble of elementary term alignment algorithms. We evaluate the performance of the system on three different industry-specific domains, where one domain was used for training and two domains were used for testing the proposed approach.

The novel approach to bilingual term alignment is the main contribution of this work. The proposed PTBA approaches are novel bilingual term alignment approaches that we have developed based on Pialign (Neubig et al., 2011), an unsupervised model for joint phrase alignment and extraction using nonparametric Bayesian methods and inversion transduction grammars. Pialign follows a similar approach to phrase table generation in statistical machine translation (SMT) (Koehn et al., 2007), however, instead of first generating word alignments and then extracting a phrase table consistent with these alignments, it joins the phases of alignment and extraction by constructing a generative model that includes phrases at many levels of granularity, from single words to full sentences.

The PTBA approaches take as input a corpus and produce a list of aligned terms as output. Specifically, the Pialign alignments are read and used for mapping that stores for each English word all the computed Slovenian alignments along with the frequency of each alignment. The same mapping is also created for the reverse direction (Slovenian to English). For each aligned sentence pair found to contain some English and Slovenian terms, we compute the matching of all English terms from this sentence against all phrases from this sentence, and the best matching is retained. The matching is computed as the ratio of the most similar substring (i.e. if the phrase contains the entire term, the result is 100%). As a result, for each English phrase found in a sentence we record which terms found in this sentence are a part of this phrase. The matching procedure is repeated also for Slovenian. Finally, for each sentence we retain only the term-to-phrase mappings that exist in both directions. That is, we store a mapping if an English term from some sentence matches an English phrase from the same sentence and a Slovenian term from the aligned Slovenian sentence matches with the aligned Slovenian phrase.

As a side result of this term-to-phrase matching procedure, we propose the following procedure to obtain a list of direct candidates for aligned terms (i.e. we identify the phrase

alignments consisting of a single term). The conditions are that the best term-to-phrase matching score is at least 95% for English and 90% (as the language is morphologically more varied) for Slovenian and the difference in length of term string and phrase string is not greater than 4[1].

The matching problem is addressed as follows: For each sentence, we have a list of phrases in English, their aligned counterparts in Slovenian, a list of terms for each English phrase and a list of terms for each Slovenian phrase. When computing the matching between English and Slovenian terms we also take into account the possibility that the terms can consist of several words. We define the matching score of a multi-word English term to a multi-word Slovenian term as the sum of best single word alignment scores among all word combinations between the terms. The matching algorithm computes the sum of all best word alignment scores.

The matching scores are accumulated for all phrases and all sentences. In the end, we obtain the probability distributions for the translation of English terms into Slovenian and Slovenian terms into English. Using this information, we can produce three translation tables: symmetric, English to Slovenian, and Slovenian to English, respectively. The symmetric table consists of only those aligned terms where the greedy probabilistic translation is the same in both directions. That is, a pair of English and Slovenian terms have each other listed as the most probable translation. The other two translation tables simply list the most likely translation in each direction. In this way, we have defined three different PTBA term alignment methods, resulting in three separate outputs of the PTBA term alignment method:

- *PTBA-1 Aligned Term List*, containing the results of the symmetric translation table.
- *PTBA-2 Aligned Term List*, containing the results of the English to Slovenian and Slovenian to English translation tables.
- *PTBA-3 Aligned Term List*, containing the list of direct alignment candidates produced as a side result of the term-to-phrase matching procedure.

To be able to effectively search the large space of various weight values, we decided to use an evolutionary algorithm to find an optimal configuration. Specifically, we utilized the genetic algorithm (GA) implementation in DEAP (Distributed Evolutionary Algorithms in Python) by De Rainville et al. (2012), an evolutionary computation framework, which can be used for rapid prototyping and testing of ideas and is designed to make algorithms explicit and data structures transparent. We start by generating a population of random sets of seven real numbers in the form of 7-tuples of weights of the 7 individual bilingual term alignment output. Each 7-tuple is used to generate a final bilingual term list and is evaluated against a database of manually annotated term pairs provided in the training dataset. We repeated the GA algorithm execution 20 times, and then calculated the average precision and standard deviation of the best performing 7-tuple of weights in each GA repetition. We selected the overall best performing 7-tuple learned on the training domain (training dataset) and tested its performance on two separate domains (test datasets).

2.3 Results

We evaluated the performance of the developed approaches in a manual evaluation setting with one annotator evaluating all results and another annotator evaluating a subset of

[1] An example, where a term matches the phrase with nearly no differences is a term upravitelj and the phrase upravitelji. As this is the only element of the phrase, we assume that the aligned phrase is the term's equivalent in English (e.g. manager)

the results to calculate the inter-annotator agreement[2]. The criterion to measure was precision of bilingual term alignment and was agreed upon with the client who provided three real-life domain-specific translation memories: automotive, financial and IT.

We measured the precision on Top N term pairs generated by individual approaches and then we investigated the performance on only top N multi-word units (MWU) which was a particular area of interest for the client. The PTBA-3 approach was in general the best performing single approach in most scenarios. By using the evolutionary algorithm to optimize for maximum precision, we were able to increase precision on the training domain (Financial) from 0.90 to 0.96. When the weights proposed by the algorithm were applied to the test domains (Automotive and IT), we achieved even higher precision (0.98 in both domains).

2.3.1 Relevance of the developed approaches

As of today, these approaches are somewhat outdated, having been surpassed by more advanced methods leveraging contextual embeddings and LLMs. Direct comparison with current state-of-the-art techniques is challenging due to the client-specific nature of the described methodology and the limited emphasis on recall, given the absence of annotated datasets. Nevertheless, the methods outlined in subsequent chapters, as well as other cutting-edge approaches, could be evaluated on these client-specific datasets and may provide superior results.

[2]The inter-annotator agreement was high with both annotators agreeing in 95% of term pairs and Cohen's kappa reaching 0.900.

TermEnsembler

An ensemble learning approach to bilingual term extraction and alignment

Andraž Repar, Vid Podpečan, Anže Vavpetič, Nada Lavrač & Senja Pollak
Jožef Stefan Institute

This paper describes TermEnsembler, a bilingual term extraction and alignment system utilizing a novel ensemble learning approach to bilingual term alignment. In the proposed system, the processing starts with monolingual term extraction from a language industry standard file type containing aligned English and Slovenian texts. The two separate term lists are then automatically aligned using an ensemble of seven bilingual alignment methods, which are first executed separately and then merged using the weights learned with an evolutionary algorithm. In the experiments, the weights were learned on one domain and tested on two other domains. When evaluated on the top 400 aligned term pairs, the precision of term alignment is over 96%, while the number of correctly aligned multi-word unit terms exceeds 30% when evaluated on the top 400 term pairs.

Keywords: bilingual terminology alignment, terminology extraction, ensemble learning, evolutionary algorithm

1. Introduction

With the onset of globalized markets, the need for effective multilingual communication has never been greater. Language industry, a term used to describe collectively the companies that offer translation and other related language services, has been steadily growing for several years and the increase in the volume of translated words brought along the need to streamline the translation process with automated solutions. In the 1990s, translation companies embraced computer-assisted translation (CAT) tools that allow them to store translations in a database and recycle them in future translation tasks.

https://doi.org/10.1075/term.00029.rep
Terminology 25:1 (2019), pp. 93–120. issn 0929-9971 | e-issn 1569-9994

Parallel to this process, another distinct (but related) development took place which revolved around terminology in the translation process. While several solutions and tools have been proposed, terminology remains one of the main problem areas for the translation industry. For example, a 2014 report[1] by SDL, a market leader in translation and terminology management software solutions, showed that among 140 companies, 51 percent of the respondents did not have a terminology management process in place, while a survey by Schmitz and Straub (2016) showed that among 800 respondents, 89.5 percent often or constantly experience that different organizational areas or employees use different terms for the same concept and that 51.9 percent of employees often or constantly cannot understand terms immediately. SDL Translation Technology Insights Series survey,[2] which focused on translation quality conducted among a mix of translation buyers, language service providers and freelance translators, found that "inconsistencies in the use of terminology" is the number one reason of translation rework (i.e. when translation is deemed not good enough and the source text has to be translated again) and recommended that, in order to improve translation quality, terminology management be prioritized.

Due to the early adoption of CAT tool technology in the translation industry, most translation companies have large repositories of translation memories. To illustrate, Gouadec (2007) reported that among more than 430 translation job advertisements surveyed, 95 percent contain a requirement for a "translation memory skill." In the period since that study, translation memories have remained a central component of any translation company business model.

This paper addresses the above-mentioned needs of the translation industry by proposing a system for semi-automated terminology extraction and alignment, currently focusing on English and Slovenian. The system, developed for one of the largest language service providers in Southeast Europe, consists of:

- A concept-oriented terminology database, where all the data is stored, allowing import from and export into industry-standard terminology management formats.
- A terminology extraction workflow, including automated extraction or import of manually defined monolingual terminology, followed by a novel approach to term alignment utilizing an evolutionary algorithm to combine the results of several individual bilingual term alignment methods.

1. SDL Research – Terminology: An End-to-End Perspective (http://www.sdl.com/download/terminology-an-endtoend-perspective/71114/). Accessed 3 March 2017.

2. Research Study 2016: Translation Technology Insights – Productivity (https://www.sdl.com/download/tti16-productivity/109572/). Accessed 3 March 2017.

– A web interface for managing the database and controlling the extraction and alignment algorithms.
– Additional functionalities for extraction of good example sentences and identification of the domain in which the term is used.

The novel approach to bilingual term alignment is the main contribution of this work. We systematically compare several existing term alignment methods, propose a novel Phrase-Table-Based Alignment (PTBA) method based on Pialign (Neubig et al. 2011), as well as a novel methodology using an evolutionary algorithm to combine solutions of an ensemble of elementary term alignment algorithms. We evaluate the performance of the system on three different domains, where one domain was used for training and two domains were used for testing the proposed approach.

This paper is structured as follows: Section 2 describes the related work, Section 3 describes the system and its methodology, Section 4 contains the experiments and results, while Section 5 contains the conclusions and plans for future work.

2. Related work

Terminology extraction refers to structuring terminological knowledge from unstructured text. Parallel translation databases (i.e. translation memories), which are omnipresent in the translation industry, lend themselves nicely to automated terminology extraction. In addition to terminology, various other types of information can be extracted, such as named entities, collocations or good examples.

In terms of input text, we can distinguish between monolingual terminology extraction, where terms are extracted from text in one language, and bilingual or multilingual terminology extraction, where the goal is to extract and align terms from text in two or more languages. A brief survey of related work is presented in Sections 2.1 and 2.2, respectively.

2.1 Monolingual term extraction

In the broadest sense, there are two different approaches to monolingual term extraction: linguistic and statistical. The linguistic approach utilizes the distinctive linguistic aspects of terms – most often their syntactic patterns, while the statistical approach takes advantage of term frequencies in the corpus. However, most state-of-the-art systems are hybrid, using a combination of the two approaches; e.g., Justeson and Katz (1995) first define part-of-speech patterns of terms and then use simple frequencies to filter the term candidates.

Many terminology extraction algorithms are based on the concepts of termhood and unithood defined by Kageura and Umino (1996). Termhood is "the degree to which a stable lexical unit is related to some domain-specific concepts" and unithood is "the degree of strength or stability of syntagmatic combinations and collocations." Termhood-based statistical measures function on a presumption that a term's relative frequency will be higher in domain-specific corpora than in the general language. Several approaches utilizing termhood have been developed, including those by Ahmad et al. (2000) and Vintar (2010). Common statistical measures are used to measure unithood, such as mutual information (Daille et al. 1994) or t-test (Wermter and Hahn 2005).

In the last few years, word embeddings – vectors of real numbers representing words on a corpus – have become a very popular natural language processing technique. The turning point was the paper by Mikolov et al. (2013) describing word2vec, a word embedding toolkit that can create vector space models much faster than previous attempts. Several attempts have already been made to utilize word embeddings for terminology extraction (e.g. Amjadian et al. (2016), Wang et al. (2016), Khan et al. (2016) and Zhang et al. (2018)).

2.2 Bilingual term extraction and alignment

At the highest level, bilingual terminology extraction can be divided into extraction from comparable and extraction from parallel corpora, where parallel corpora are composed of source texts and their translations in one or more different languages, while comparable corpora are composed of monolingual texts collected from different languages using similar sampling techniques (McEnery et al. 2006). For alignment of terms between the two languages, the methods typically utilize the idea that a term and its translation tend to occur in similar lexical contexts (Daille and Morin 2005).

In the language-industry context, taking into account parallel bilingual sentence pairs, stored in the translation memory, brings significant advantages to the task of terminology extraction. Broadly speaking, there are two distinct approaches to bilingual terminology extraction from parallel corpora according to Foo (2012):

- Align-extract, where we first align single and multi-word units in parallel sentences and then extract the relevant terminology from a list of candidate term pairs, and
- Extract-align, where we first extract monolingual candidate terms from both sides of the corpus and then align the terms.

A state-of-the-art align-extract approach is proposed by Macken et al. (2013) utilizing a chunk-based alignment method to produce a list of candidate term pairs, which are then filtered using statistical methods.

The extract-align approach is the more common of the two. Kupiec (1993) describes an algorithm for noun phrase extraction followed by alignment with a statistical estimation algorithm, achieving precision of 90 percent on the highest ranking candidate pairs. Vintar (2010) describes an extract-align approach named "bag-of-equivalents", where after monolingual extraction, the term pairs are aligned with the help of word alignment probabilities. Baisa et al. (2015) describe a frequency-based term alignment algorithm utilizing a variation of logDice to score the strength of the candidate term pair alignment. Haque et al. (2014) first generate candidate terms monolingually and then build a phrase table using the Moses toolkit (Koehn et al. 2007) and compare the extracted terms with the phrases in the table. Precision among the top 100 candidate term pairs often exceeds 90 percent. Aker et al. (2013) treat bilingual term alignment as a binary classification task, achieving good results. More recently, Hazem and Morin (2017) experiment with word embeddings used to augment bilingual terminology extraction from specialized comparable corpora (achieving precision of 70.9 percent).

The approach proposed in this paper is based on the idea of utilizing evolutionary algorithms which mimic biological evolution (i.e. reproduction, mutation, selection) to optimize the stated objective. Specifically, we use the genetic algorithm implementation in DEAP (*Distributed Evolutionary Algorithms in Python*) by Fortin et al. (2012) to build a term alignment ensemble.

3. TermEnsembler system and methodology

In this section, we describe the functionality of the developed TermEnsembler system, starting with the system overview and the background technologies used, and then focusing on bilingual term alignment as the main contribution of this paper.

3.1 System overview

The TermEnsembler system extracts bilingual terminology from English and Slovenian texts, and stores it into a concept-based terminology database, meaning that the entries are organized to correspond to a concept (cf. the general theory of terminology proposed by Wüster (1979)), but a concept might have more than one corresponding designator. It is a semi-automated system, meaning that the user can select several extraction parameters and manually curate the monolingual

extraction results for better bilingual alignment. While the system currently supports two languages (English and Slovenian), additional languages can be added by implementing appropriate language-specific background technologies similar to the ones described in this paper. In addition to the extraction of individual terms in each of the two languages (extracted using the approach described in Section 3.2), it also stores aligned term pairs (aligned using the approach described in Section 3.3). We have also developed a method for extracting good examples and domains, but as these are additional functionalities, we refer the reader to the previous papers by Repar and Pollak (2017a, 2017b).

The system relies on several background resources and technologies, used in different components of the system:

- *Preprocessing:* Texts are extracted from the translation memory (TMX) and preprocessed using the part-of-speech tagger and Wordnet lemmatizer from NLTK (Bird et al. 2009) for English and using the ReLDI tagger and lemmatizer (Ljubešić and Erjavec 2016) for Slovenian.
- *Monolingual term extraction:* Monolingual term extraction method LUIZ-CF by Pollak et al. (2012), extending LUIZ (Vintar 2010), is used as a basis for our upgraded LUIZ-CF++ term extraction approach.
- *Bilingual term alignment:* We use the Pialign phrase table extraction functionality (Neubig et al. 2011) as a basis for implementing three different bilingual term alignment approaches PTBA-1, PTBA-2 and PTBA-3 used in our experiments. In the reimplementation of bilingual LUIZ, we use Giza++ for word alignment (Och and Ney 2003). For weight assignment in our ensemble approach, we use the evolutionary computation framework DEAP (Distributed Evolutionary Algorithms in Python) by Fortin et al. (2012).

The overall structure of the system is shown in Figure 1. The starting point is a bilingual corpus in the standard translation memory format TMX, from which also available metadata, such as term domain or language variety can be extracted. The text is extracted and preprocessed resulting in a list of aligned lemmatized and POS-tagged sentence pairs. These pairs are sent into the additional metadata extraction (e.g., when domain information is not available in the TMX) and the monolingual extraction process, which results in two separate monolingual term lists (for TL1 and TL2). At this point, these two term lists can be curated by the user of the system. The (raw or curated) term lists are then taken as input to the bilingual alignment process (described in detail in Figure 2), which produces the final list of aligned term pairs. Finally, these term pairs are entered in the termbase alongside the metadata extracted in the step described above.

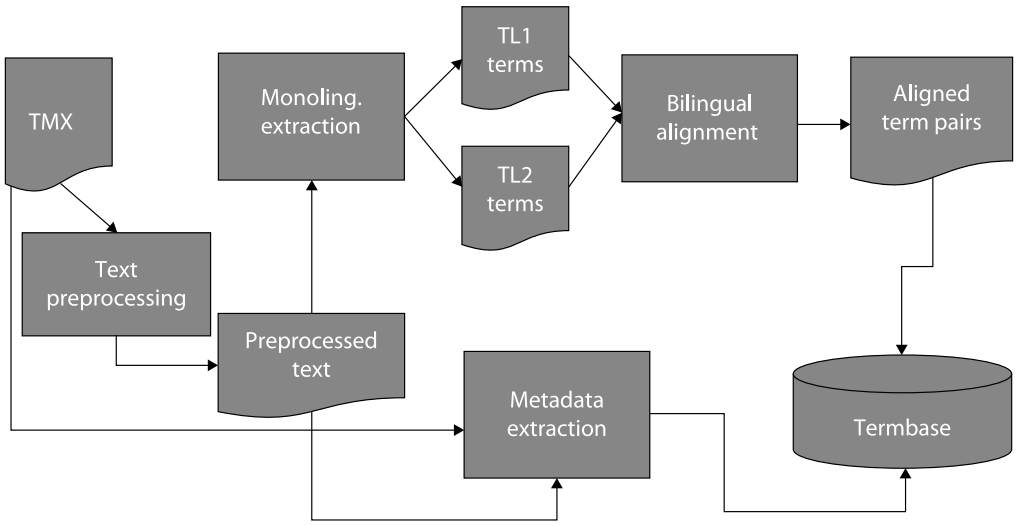

Figure 1. TermEnsembler: Methodology and components of the TermEnsembler system. Note that at several points human curation is possible (after monolingual extraction, after bilingual alignment or when accepting terms and metadata in the termbase. The monolingual step can also be skipped if the monolingual term lists are manually provided

3.2 Monolingual term extraction: LUIZ-CF++ upgrade of LUIZ-CF

The implemented monolingual term extraction approach LUIZ CF++ is based on the LUIZ hybrid approach by Vintar (2010) and refined with scoring and ranking functions implemented in LUIZ-CF by Pollak et al. (2012). The LUIZ approach is based on a list of part-of-speech patterns and a formula for comparison of term frequency between a domain corpus and a general language corpus (we used frequency lists from corpus Kres (Logar et al. 2012) for Slovenian and the British National Corpus (2007) for English).

In LUIZ-CF++, used in our experiments, we upgraded the LUIZ-CF monolingual term extraction approach by implementing the following additional functionalities:

- *Near-duplicates detection*: When importing the terms, the near duplicates (e.g. the orthography with or without spaces or hyphens, British and American English spellings) are detected and not created as new entries, but can be added as term variants of existing entries.
- *Nested term filtering*: According to Frantzi et al. (2000), nested terms are the terms that appear within other longer terms, and may or may not appear by themselves in the corpus. If the difference between a term and its nested term is below a certain threshold (which, in our case, can be defined by the user), only the longer term is returned. If not, both terms are included in the final output.

3.3 Bilingual term alignment: A novel ensemble learning approach

In this section, we describe the core part of TermEnsembler, i.e. the bilingual term alignment methodology implementing the *extract-align* approach explained in Section 2.2. Having implemented seven elementary term alignment approaches (3 existing, one modified, and 3 novel variants based on Pialign), this section introduces a novel ensemble-based approach combining the selected elementary term alignment approaches using an evolutionary algorithm.

We start by a brief outline of the proposed term alignment approach, illustrated in Figure 2. The input to the proposed TermEnsembler's bilingual term alignment methodology are two term lists (TL1 and TL2), which are automatically extracted using the monolingual extraction component (described in Section 3.2) or are human-defined. These two term lists are fed into seven individual bilingual term alignment algorithms that produce a total of 7 separate lists of aligned term pairs (*aligned term lists* or ATL), ranked by their alignment probability score as described in Section 3.3.1. The outputs of each alignment method are first normalized (separately) to the [0,1] interval, then fed into the evolutionary weights optimization algorithm described in Section 3.3.3 (which uses an external *ground truth list* (GTL) of manually annotated term pairs) to produce an optimal set of weights. These weights are then used to merge the seven ATLs into the final merged ATL using the procedure from Section 3.3.2.

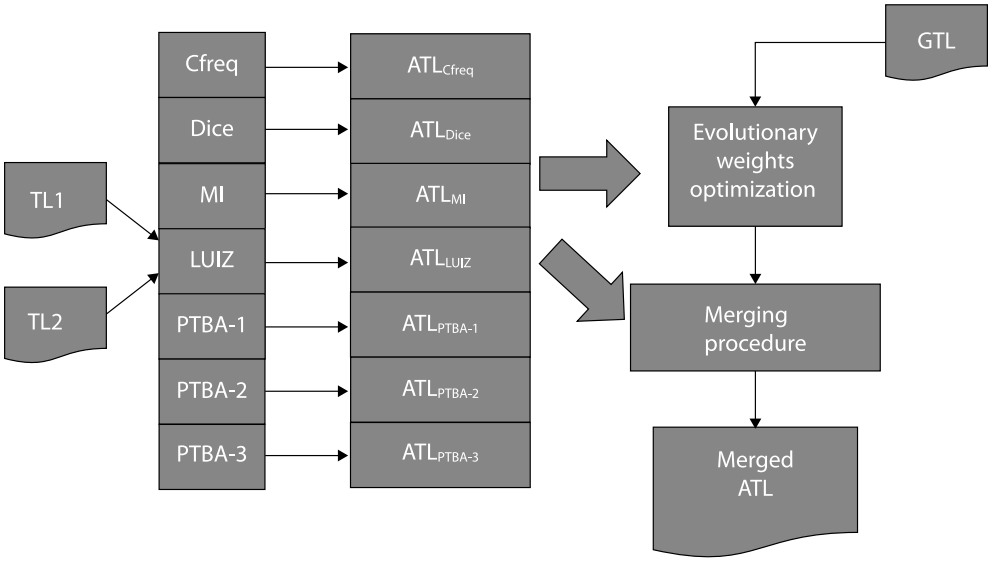

Figure 2. TermEnsembler's bilingual term alignment methodology

3.3.1 *Individual bilingual term alignment algorithms*

Each term alignment component described in this section produces a list of aligned term pairs ranked by their alignment scores, which are normalized between 0 and 1. The calculation of the scores is described below. The first four reimplemented approaches produce each one output (one aligned term list), while the last, novel approach, has three variants, leading to a total of seven output lists of aligned term pairs.

Co-frequency

Co-frequency *cofreq(t_S, t_T)* simply counts the number of sentences in which a term (t_S) from a source language S and a term (t_T) from target language T co-occur in the same sentence pair. The higher the co-occurrence count, the higher the probability that the terms are a correct term pair. This is the simplest of the used approaches and is completely language independent, but it does not take into account any language specifics. Because of that, it also requires a larger input corpus to produce sensible results.

Dice

This approach to bilingual terminology extraction is based on the Dice algorithm (Dice 1945). The co-frequency score from the previous component is used in the calculation of the Dice score, defined as follows:

$$dice(t_S,t_T) = 2\,\frac{cofreq(t_S,t_T)}{freq(t_S) + freq(t_T)}$$

where (t_S) and (t_T) are source and target terms, respectively. The *freq(t)* function stands for the frequency of term t in the entire corpus. A score based on Dice is used also in Sketch Engine (Baisa et al. 2015).

Mutual information

Similar as Dice, MI (Church and Hanks 1990) calculates term alignment by taking into account the co-frequency of source and target terms and the individual frequency of each term. It is defined as follows:

$$MI(t_S,t_T) = \log_2 \frac{cofreq(t_S,t_T)}{freq(t_S)\,freq(t_T)}$$

It usually contains the multiplication with N (in our case the number of candidate terms), but since in our case N is constant across terms, we can omit it if we just want to rank the terms.

BI-LUIZ+

We used a modified version of the bilingual component of the LUIZ approach, described by Vintar (2010). This approach takes as input two lists of term candidates (one for the source language and one for the target language) and word alignment pairs (with probabilities). The original paper uses the Twente aligner (Hiemstra 1998), while we used the GIZA++ (Och and Ney 2003).[3]

Using the alignments, the best matches (1 or more) are computed for each source term as follows: given a source term, we iterate through all target terms. For each target term we compute a score by summing the probabilities that a target token is a translation of a source token. Note that in the original paper by Vintar (2010) the equivalence score takes all single-word probabilities and divides them by the number of words, but dividing is not performed in our re-implementation as in the testing phase it produced worse results.[4] If the score is non-zero, we add the target term to the list of candidates.

Novel Phrase-Table-Based Alignment (PTBA) approaches PTBA-1, PTBA-2 and PTBA-3

The proposed PTBA approaches are novel bilingual term alignment approaches that we have developed based on Pialign (Neubig et al. 2011), an unsupervised model for joint phrase alignment and extraction using nonparametric Bayesian methods and inversion transduction grammars. Pialign follows a similar approach to phrase table generation in statistical machine translation (SMT) (Koehn et al. 2007), however, instead of first generating word alignments and then extracting a phrase table consistent with these alignments, it joins the phases of alignment and extraction by constructing a generative model that includes phrases at many levels of granularity, from single words to full sentences. Similar to Haque et al. (2014), the PTBA approach uses machine translation phrase tables for term alignment, but differs from it in several aspects described below.

The proposed PTBA approach takes as input a corpus and produces the list of aligned terms as output. Specifically, the Pialign alignments are read and used for mapping that stores for each English word all the computed Slovenian alignments along with the frequency of each alignment. As illustration, take the following example:

manager → upravitelj (20%), upravljavec (30%), upravljavec premoženja (50%)

The same mapping is also created for the reverse direction (Slovenian to English). For each aligned sentence pair found to contain some English and Slovenian terms, we compute the matching of all English terms from this sentence against

3. We had to use a different alignment method since the Twente aligner does not work anymore.

4. In communication with Vintar it has been confirmed that division has been later excluded from the formula.

all phrases from this sentence, and the best matching is retained. The matching is computed as the ratio of the most similar substring (i.e. if the phrase contains the entire term, the result is 100%). As a result, for each English phrase found in a sentence we record which terms found in this sentence are a part of this phrase. The matching procedure is repeated also for Slovenian. Finally, for each sentence we retain only the term-to-phrase mappings that exist in both directions. That is, we store a mapping if an English term from some sentence matches an English phrase from the same sentence and a Slovenian term from the aligned Slovenian sentence matches with the aligned Slovenian phrase.

As a side result of this term-to-phrase matching procedure, we propose the following procedure to obtain a list of direct candidates for aligned terms (i.e. we identify the phrase alignments consisting of a single term). The conditions are that the best term-to-phrase matching score is at least 95% for English and 90% (as the language is morphologically more varied) for Slovene and the difference in length of term string and phrase string is not greater than 4. An example, where a term matches the phrase with nearly no differences is a term *upravitelj* and the phrase *upravitelji*. As this is the only element of the phrase, we assume that the aligned phrase is the term's equivalent in English (e.g. *manager*).

The matching problem is addressed as follows: For each sentence, we have a list of phrases in English, their aligned counterparts in Slovenian, a list of terms for each English phrase and a list of terms for each Slovenian phrase. When computing the matching between English and Slovenian terms we also take into account the possibility that the terms can consist of several words.

We define the matching score of a multi-word English term to a multi-word Slovenian term as the sum of best single word alignment scores among all word combinations between the terms. Consider the following example:

English sentence	*The name of the share class Allianz….*
Slovenian sentence	*Ime razreda delnic Allianz…*
English phrase	*The name of the share class*
Slovenian phrase	*Ime razreda delnic*
English terms	*share, share class*
Slovenian terms	*delnica, razred delnic*

The matching algorithm computes the sum of all best word alignment scores. For example *score(share, delnica) + score(class, razred)* is the alignment score for terms *share class* and *razred delnic* (the word (mis)alignments *share-razred* and *class-delnica* have very low or possibly zero scores and are not added to the sum).

The matching scores are accumulated for all phrases and all sentences. In the end, we obtain the probability distributions for the translation of English terms into Slovenian and Slovenian terms into English. Using this information, we can produce three translation tables: *symmetric, English to Slovenian, and Slovenian to English*, respectively. The symmetric table consists of only those aligned terms

where the greedy probabilistic translation is the same in both directions. That is, a pair of English and Slovenian terms have each other listed as the most probable translation. The other two translation tables simply list the most likely translation in each direction. In this way, we have defined three different PTBA term alignment methods, resulting in three separate outputs of the PTBA term alignment method:

- *PTBA-1 Aligned Term list*, containing the results of the symmetric translation table.
- *PTBA-2 Aligned Term list,* containing the results of the English to Slovenian and Slovenian to English translation tables.
- *PTBA-3 Aligned Term list,* containing the list of direct alignment candidates produced as a side result of the term-to-phrase matching procedure.

3.3.2 *Final term pair ranking by ensemble-based weighting of separate lists of term pairs*

This section presents the key part of the developed methodology for ranking of aligned term pairs, i.e. the mechanism for assigning weights to separate lists of term pairs obtained by individual term alignment algorithms, and the merging mechanism using an ensemble weighting approach.

The ensemble score (Escore) is computed from two separate weighting scores:

- the algorithm weight (w), and
- the term pair score (score), normalized to [0,1].

A merging procedure for computing the final ensemble score Escore takes the individual term pair scores (score) from each of the seven elementary algorithms, together with weights for each approach provided by the user or assigned by automated means (i.e. the evolutionary algorithm approach explained below) and returns the final aligned term list, re-normalized on the [0,1] interval.

Merging procedure

1. For all **term pairs** (t_S, t_T) **compute** $Escore(t_S, t_T)$:

$$
\begin{aligned}
Escore(t_S, t_T) = {} & w_{cofreq}\, score_{cofreq}(t_S, t_T) + \\
& w_{dice} \cdot score_{dice}(t_S, t_T) + \\
& w_{mi} \cdot score_{mi}(t_S, t_T) + \\
& w_{mi} \cdot score_{mi}(t_S, t_T) + \\
& w_{luiz} \cdot score_{luiz}(t_S, t_T) + \\
& w_{PBA1} \cdot score_{PTBA\text{-}1}(t_S, t_T) + \\
& w_{PBA2} \cdot score_{PTBA\text{-}2}(t_S, t_T) + \\
& w_{PBA3} \cdot score_{PTBA\text{-}3}(t_S, t_T)
\end{aligned}
$$

2. Compute *Normalized Escore*(t_S, t_T)∈[0,1]
3. Rank *term pairs* (t_S, t_T) in decreasing order of their *Normalized Escore*(t_S, t_T)

3.3.3 *Evolutionary weighting of term alignment algorithms*

To be able to effectively search the large space of various weight values, we decided to use an evolutionary algorithm to find an optimal configuration. Specifically, we utilized the genetic algorithm (GA) implementation in DEAP (*Distributed Evolutionary Algorithms in Python*) by Fortin et al. (2012), an evolutionary computation framework, which can be used for rapid prototyping and testing of ideas and is designed to make algorithms explicit and data structures transparent. The GA algorithm starts with a random population and then applies crossover (producing new (children) members of the population from existing (parent) members) and mutation (randomly changing individual members – similar to biological mutation) operations for a successive number of generations. In each generation, the children are evaluated using a custom evaluation function and those that perform better than the parents are retained, while those that perform worse are discarded which eventually leads to an optimal result.

We start by generating a population of random sets of seven real numbers in the form of 7-tuples of weights of the 7 individual bilingual term alignment outputs:

$$(w_{cofreq}, w_{dice}, w_{mi}, w_{luiz}, w_{PBA1}, w_{PBA2}, w_{PBA3})$$

Each 7-tuple is used to generate a final bilingual term list (see Section 3.3.2) and is evaluated against a database of manually annotated term pairs provided in the training dataset. We used the parameters suggested in the DEAP documentation: number of generations: 100; population: 100; crossover probability: 0.5; mutation probability: 0.2.

We repeated the GA algorithm execution 20 times, and then calculated the average precision and standard deviation of the best performing 7-tuple of weights in each GA repetition. We selected the overall best performing 7-tuple learned on the training domain (training dataset) and tested its performance on two separate domains (test datasets). DEAP can be set up to optimize a single objective (i.e. precision among the Top 400 term pairs as in Section 4.4.1) or multiple objectives (i.e. precision among the Top 400 term pairs and number of correct *multi-word unit* (MWU) term pairs as in Section 4.4.2) at the same time.

4. Experiments and results

This section describes the experiments conducted to evaluate the TermEnsembler bilingual term alignment methodology and the datasets used in the experiments, followed by the results of the experiments and a qualitative analysis of errors.

4.1 Experimental setting

In these experiments, our goal was to find the best weight configuration for the 7 outputs produced by the individual term alignment components. To do so, we first evaluated the outputs individually in terms of overall precision and precision of MWU (*multi-word unit*) terms and then tried to find the best weight configuration using the evolutionary algorithm. We learned the best weight configuration on one domain (*Financial*) and then tested it on two others, non-related domains (*IT* and *Automotive*), by which we show that it is applicable to different domains.

The experimental setting was as follows. In creating the monolingual term lists as described in Section 3.2, we included only the terms that appear more than 10 times in the dataset.

– The evaluation criterion was the precision of term alignment, where the criterion for annotation was proper alignment, and not whether the individual English and Slovenian units are actually terms or not.

The latter requires further clarification.

– As bilingual term alignment is the main focus of this paper, we were primarily concerned with whether the terms are aligned properly (whether the terms are translation equivalents) and not whether the terms are true terms in each language.[5] For illustration, consider the following two examples:

  exchange rate – menjalni tečaj
  end of march – konec marca

  In the first example, both terms (English and Slovenian) are true terms according to the definition of a term from ISO 1087 ("verbal designation of a general concept in a specific subject field"), while the terms in the second example are much less likely to be considered terms in the sense of ISO 1087. However, for the purposes of evaluating the bilingual alignment algorithm both examples were considered correct.

5. An evaluation by a subject-matter expert reviewing the top 200 term pairs produced by the system showed that 74.5% of them are true terms.

– The evaluation was performed by a single annotator, which is the only realistic setting in a language-industry environment. Nevertheless, for inter-annotator evaluation, we acquired a second annotator to annotate a subset of the final output produced (and previously annotated by the main annotator) with the final weight configuration (see Section 4.4) on the Financial domain. The inter-annotator agreement was high, with both annotators agreeing in more than 95% of term pairs and Cohen's kappa (Cohen 1968) reaching 0.900. This denotes almost perfect agreement according to Landis and Koch (1977), and we can safely assume that annotations performed by a single annotator are highly accurate.

Note that in addition to measuring the precision of term alignment, we initially also considered measuring the recall, for which we would need a dataset containing manually annotated term pairs. However, measuring recall proved to be practically less relevant. The client arrived at the conclusion that in a production environment of a language service provider, the recall is not of particular importance, while it is much more important that term extraction output be precise, requiring no or minimal further processing or manual selection. As will be shown in Section 4.4, TermEnsembler produces a large number of correct term pairs, which satisfies the needs of the client. However, for the purpose of this article, we did evaluate the recall on a small gold standard term list in Section 4.4.3.

4.2 Data

In our experiments we used three distinct datasets, all coming from a production environment of a language service provider.

– Financial. This translation memory contains segments from a long-term translation project in the financial domain, specifically annual reports of investment funds and various related documentation. It has 18,197 segments (i.e. bilingual segment pairs) with 396,295 words in English and 354,862 words in Slovenian. The default configuration of the monolingual extractor returned 1,723 English and 1,953 Slovenian terms. This dataset was used to find the best weight configuration with the evolutionary algorithm.
– IT. This translation memory was used in a long-term software localization project. Most segments contain user interface strings and a smaller portion also contains user assistance (i.e. help articles) content. It has 40,599 segments (i.e. bilingual segment pairs) with 523,819 words in English and 473,430 words in Slovenian. The default configuration of the monolingual extractor returned 2,234 English and 2,477 Slovenian terms. This dataset was used to test the best weight configuration found with the evolutionary algorithm on the Financial dataset.

– Automotive. This translation memory was used in a long-term project for a customer from the automotive industry and contains segments from user manuals, internal service documentation and customer-facing promotional materials. It has 65,516 segments (i.e. bilingual segment pairs) with 861,665 words in English and 779,145 words in Slovenian. The default configuration of the monolingual extractor returned 3,122 English and 3,879 Slovenian terms. This dataset was used to test the best weight configuration found with the evolutionary algorithm on the Financial dataset.

Detailed statistics for each dataset, including the number of terms obtained by monolingual terminology extraction, are presented in Table 1.

Table 1. Detailed statistics of the three datasets used in the experiments

	Financial	IT	Automotive
Total segments	18,197	40,599	65,516
Total English words	396,295	523,819	861,665
Total Slovenian words	354,862	473,430	779,145
Unique English words	11,365	21,711	25,591
Unique Slovenian words	20,093	31,973	43,406
English terms	1,723	2,234	3,122
Slovenian terms	1,953	2,477	3,879

4.3 Experimental comparison of individual bilingual term alignment components

In this section, we systematically compare the performance of individual bilingual term alignment components from two aspects. First, we focus on the overall precision of the Top N term pairs produced by each component, and then we turn our attention to MWU (*multi-word unit*) term pairs found in the top N term pairs produced by the individual components.

4.3.1 *Precision of individual term alignment components*

Table 2 provides the results for precision for each method on the Financial dataset. We can observe that two PTBA methods have the highest precision, followed by another PTBA method and the three frequency-based components (Co-frequency, Dice and Mutual information), while BI-LUIZ+ has the lowest precision.

Table 2. Precision of individual bilingual alignment components on the Financial dataset on the Top 100, Top 200, Top 400 and Top 800 term pairs according to their (normalized) alignment score

	Total term pairs	Top 100		Top 200		Top 400		Top 800/ Total	
		Corr.	Prec.	Corr.	Prec.	Corr.	Prec.	Corr.	Prec.
Co-freq	1,492	60	0.600	111	0.555	175	0.438	292	0.366
Dice	1,492	57	0.570	128	0.640	272	0.680	511	0.693
MI	1,492	59	0.590	120	0.600	229	0.573	398	0.498
BI-LUIZ+	1,561	43	0.430	82	0.410	136	0.340	228	0.285
PTBA-1	591	93	0.930	183	0.915	350	0.875	**486**	**0.822**
PTBA-2	1,341	74	0.740	148	0.740	246	0.616	436	0.546
PTBA-3	674	**98**	**0.980**	**193**	**0.965**	**360**	**0.900**	523	0.777

Note that since the total number of term pairs of PTBA-3 and PTBA-1 is lower than 800, the last column denotes precision on the total number of pairs (i.e. Top 674 and Top 591, respectively).

4.3.2 *Single vs. multi-word unit terms*

While precision is the most important performance indicator of a bilingual term alignment algorithm, we also wanted to have more details on the ratio between single and multi-word terms in the outputs, because the client communicated that having translations of multi-words terms is much more useful than just simple one-word units. Since we are looking at bilingual term pairs, we consider a pair to be a single-word unit if both terms (English and Slovenian) are single-word units, and multi-word if at least one of the terms is a multi-word unit (MWU). For illustration, see the three examples below:

issuance – izdaja *SINGLE-WORD UNIT*
registrar – agent za registracijo *MULTI-WORD UNIT*
stock market – borzni trg *MULTI-WORD UNIT*

Specifically, we looked at how many of the top N terms produced by individual components are correct MWU term pairs. This decision was again reached in communication with the client who wanted to have the ability to request a specific number (N) of term pairs to be returned by TermEnsembler and our goal was to make the returned term pairs as good as possible, both in terms of overall precision and in the number of correct MWU terms.

In Table 3, we can observe that the Dice algorithm produces the most correct term pairs in all 4 scenarios, closely followed by MI. BI-LUIZ+ produces a lot of multi-word terms but its precision (calculated as correct MWU terms divided by all MWU terms in the top N term pairs) is relatively low, while the PTBA methods

produce few MWU term pairs in the Top 100 pairs, but improve in this respect in Top 200, Top 400 and Top 800 scenarios.

Table 3. Total number of MWU term pairs (and their precision) in top N terms, correct MWU term pairs on the Financial dataset

	Top 100		Top 200		Top 400		Top 800/Total	
	Cor/tot	**Prec**	**Cor/tot**	**Prec**	**Cor/tot**	**Prec**	**Cor/tot**	**Prec**
Co-freq	2/21	0.420	7/49	0.143	17/128	0.133	49/383	0.128
Dice	**52**/94	0.553	**106**/175	0.606	**198**/320	0.619	**358**/589	0.608
MI	50/87	0.575	102/178	0.573	187/351	0.533	295/678	0.435
BI-LUIZ+	43/100	0.430	82/200	0.410	103/363	0.284	136/680	0.200
PTBA-1	20/24	0.833	51/61	0.836	133/170	0.782	199/273	0.729
PTBA-2	15/38	0.395	39/85	0.459	90/234	0.385	194/527	0.368
PTBA-3	14/14	**1.000**	54/57	**0.947**	130/146	**0.890**	218/278	**0.784**

Note that since the total number of term pairs of PTBA-3 and PTBA-1 is lower than 800, the last column denotes precision on the total number of pairs (i.e. Top 674 and Top 591, respectively).

4.4 Results of the TermEnsembler's bilingual term alignment approach

The key question in our system is how to determine the optimal configuration of weights for the merging script described in Section 3.3. Table 2 and Table 3 above clearly show that some of the methods are much more effective than the others. Similar to the reasoning in Section 4.3, we want to test two distinct scenarios:

- In the first one, we want to find the best overall precision.
- In the second one, we want to find the best compromise between the overall precision and the number of correct multi-word units.

We decided to focus the evaluation of the weight configuration on the top 400 term pairs, because the client believes that 400 terms are enough to produce a useful terminological resource in a standard translation project. In other words, we try to optimize the configuration to return the best results on the top 400 term pairs. Also, the starting point for comparison is the result of the PTBA-3 component that has an overall precision of 0.900 and returns 130 correct multi-word unit term pairs (see Table 2). This means that any weight configuration would need to improve on these results.

As evident from Table 4, assigning the same weight to all components does not yield results superior to the PTBA-3 component. The same is true if we assign weights according to their individual precision (calculated in Table 2) relative to the lowest value (i.e. the weight of BI-LUIZ+ is 1.0 and the rest are calculated

proportionally). This is why we decided to use the DEAP evolutionary algorithm described in Section 3.3 for weight configuration.

4.4.1 *Optimizing for optimal precision*

In the first experiment, we wanted to construct a weight configuration that would result in the highest possible precision, which means that we minimize the number of incorrect pairs. We performed 20 repetitions of the evolutionary algorithm execution. The average precision of the best performing 7-tuples of weights in each of the 20 repetitions was 0.949 with a standard deviation of 0.009. The overall best precision of 0.960 was achieved by three different weight configurations (see Table 5),[6] showing that the evolutionary algorithm exceeds the results of PTBA-3 by 6% (see Table 4).

Table 4. Results of the various weight configurations on the Financial domain

	Top 400
PTBA-3	0.900
Equal weights	0.725
Precision weights	0.732
Evolutionary algorithm	**0.960**

To test whether this configuration can be applied universally, we used it to evaluate precision on two additional domains: *Automotive* and *IT*. To do so, we tested all three configurations from Table 5 and calculated the average overall precision. As can be observed from Table 6, the weight configuration produced by the evo-

6. The calculated weights show that the PTBA-3 component is always the most significant one, followed by PTBA-1, and next Cofreq followed by all other methods (which can in some cases even have negative weights). Several factors that contribute to the actual magnitude of weights have to be taken into account when interpreting the results. First, the weights are computed using different heuristics. Second, the components produce results of different lengths and those returning a small number of mostly correct results are likely to obtain a higher weight. Next, the evolutionary algorithm will try to adjust the weights in such way that segments of high ranked correct results will make it to the final list. If the same or similar segment of correct results appears at the bottom of the list of another component, its promotion to the final list is likely to be too costly as this would also promote several incorrect results. For example, the reason for the negative weights in some of the repetitions in Table 5 is that the scores assigned by a particular component (i.e. PTBA-2) are too high compared to other components. This is confirmed by the results of the manual evaluation of individual components in Table 2 where we can observe that PTBA-2 has a significantly lower precision than PTBA-1 or PTBA-3. The weights of the remaining 4 components are significantly lower, close to 0, with the highest one of them being Cofreq.

Table 5. The best performing weight configurations when optimizing overall precision using an evolutionary algorithm

Rep #	Cofreq	Dice	MI	Luiz	PTBA-1	PTBA-2	PTBA-3
3	0.619	0.196	0.010	0.053	4.481	−2.867	11.046
8	0.327	0.086	0.008	0.022	1.564	0.137	5.494
10	0.561	0.106	−0.017	0.104	2.177	−0.758	10.268

lutionary algorithm returns good results on unseen data (*IT* and *Automotive*) as well, with precision on unseen data actually exceeding the precision on the training data (i.e. *Financial* domain).

Table 6. Precision of the weight configuration produced by the evolutionary algorithm on the Financial domain and applied to the Automotive and IT domain. The results were obtained as an average precision of the three weight configurations shown in Table 5

	Top 400
Financial	0.960±0.000
Automotive	0.984±0.001
IT	0.984±0.001

4.4.2 *Optimizing for a compromise between optimal precision and number of correct multi-word unit term pairs*

In the next step, we modified the evolutionary algorithm to optimize the configuration for the highest precision and the largest number of multi-word units at the same time. While the equal weight configuration and the weight configuration based on individual precision values produce a higher number of MWUs, they also introduce a fair amount of noise resulting in lower precision. As is evident from Table 7, the configuration produced by the evolutionary algorithm has the highest precision while maintaining a decent amount of MWUs (a high number of which are also correct – MWU precision of 0.919). The results closest to this configuration are returned by the PTBA-3 component, but the number of MWUs is significantly lower.

These results were achieved by running 20 repetitions of the evolutionary algorithm and selecting the best weight configuration based on the following criterion: the best configuration has the highest number of correct MWUs and must have an overall precision greater than the best individual component (in our case, PTBA-3). The best weight configuration was thus produced in repetition 19 and had the weights shown in Table 8.

Table 7. Overall precision, total number of MWUs, number of correct MWUs and precision of MWUs of the configuration produced by the evolutionary algorithm compared to various other configurations, measured on the Financial domain

	Precision	Total MWUs	Correct MWUs	MWU precision
PTBA-3	0.900	146	130	**0.890**
Equal weights	0.725	311	205	0.659
Precision weights	0.733	**312**	**208**	0.667
Evolutionary algorithm	**0.955**	185	170	0.919

Table 8. The best performing weight configuration when optimizing for a compromise between optimal precision and number of correct multi-word unit term pairs

Rep #	Cofreq	Dice	MI	Luiz	PTBA-1	PTBA-2	PTBA-3
19	0.219	0.229	0.009	0.116	2.855	−4.739	11.470

Once again, we tested whether the configuration produced by the evolutionary algorithm can be used universally by applying it to two additional domains: Automotive and IT. The results can be found in Table 9.

Table 9. Top 400 results of the weight configuration produced by the evolutionary algorithm on the Financial domain and applied to the Automotive and IT domain

	Precision	Total MWUs	Correct MWUs	MWU precision
Financial	0.955	**185**	**170**	0.919
Automotive	**0.990**	153	151	**0.987**
IT	0.985	130	126	0.969

In both domains, the results are similar to what we observed in the *Financial* domain. In fact, the results are even better in the two new domains with overall precision in the Top 400 term pair candidates exceeding 98%, and the MWU precision above 96%. The actual ratio of correct MWU terms among the Top 400 terms is 38% on the *Automotive* domain and 32% on the *IT* domain. We decided to use this configuration as the final configuration in the client's production environment.

4.4.3 *Recall of the TermEnsembler system*

Due to the client's preference, the majority of our experiments were focused on precision, but we did also evaluate recall on a corpus subsample, where a gold standard termlist of 88 financial terms was produced by manual expert annotation. With the final weight configuration (used in the production environment) the recall of the TermEnsembler system was 60%.

4.5 Qualitative analysis of errors

To better understand the types of errors that the system makes, for each of the three domains we have performed a qualitative analysis of the first 50 incorrect term pairs[7] among the list of 800 top ranked term pairs suggested by the system, using the final weight configuration suggested by the evolutionary algorithm. We observed that most of the errors are due to discrepancies between the English and Slovenian monolingual extraction process, rather than due to the incorrect alignment procedure, and that many incorrect term pairs can be considered "partially correct". We illustrate several examples of incorrect alignments below, starting with minor errors followed by some more severe cases of misaligned terms.

In some of the highly ranked term pairs, one part of a term in one language is missing because the term was incorrectly extracted, which results in partially correct term pair, such as (the word in brackets was not extracted):

Financial: *interest (rate) – obrestna mera*
Automotive: *(quick) repair kit – komplet za hitro popravilo*
IT: *missing (value) – manjkajoča vrednost*

A particularly difficult issue for the system are product names. Because they may not follow standard language rules regarding the construction of terms, they are difficult to detect without a pre-defined product name list or a well performing named entity recognition system. Consequently, many of the incorrectly extracted named entities contain parts of product names. The Financial dataset in particular has a high number of named entities, which is a reason for lower results compared to the other two corpora. Such examples include:

Equity – delnica[8]
BNP Paribas – Paribas
Flexible Bond Strategy – Bond Strategy

In a limited number of cases, the monolingual terms and the alignment itself are correct, but the resulting term pair is not correct. In the two examples from the Automotive dataset, the source text uses *miles per gallon* to denote gas mileage, but the Slovenian translation (due to the preferences of the customer) uses *kilometers per 100 liters*. A similar case can be observed with units denoting weight.

Mile – km
Lb – kg

7. The positions of the 50th incorrect term pair for all three domains: 518 for Financial, 756 for Automotive, and 661 for IT.

8. Note that "equity" can appear either as a common noun (i.e. equity = assets) or as a part of a proper noun (e.g., Global Equity Climate Change).

In a smaller number of cases close to the bottom of the list of extracted term pairs, the alignment is completely off and the meaning of the source term is not the same as the meaning of the target term (which can be explained by the frequent co-occurrence of the terms in the text), for example:

Financial: *gross national income – svetovna banka*
Automotive: *similar heavy object – pritrjen nosilec koles*
IT: *folder number – znesek kredita*

Finally, we compared the ratio between the two major error types in the three domains (see Table 10). In the *Financial* and *Automotive* domains, the majority of the incorrect terms can be ascribed to the category "Partially correct", which are predominantly errors arising from incorrect monolingual extraction (but could also be related to incorrect translation or wrong alignment of the two terms). Because the monolingual term is missing a word or several words or contains redundant words, the resulting term pair was not classified as correct. However, the alignment is not completely wrong nor completely useless, because the term can be quickly corrected in a semi-automated terminology setting.

Table 10. A comparison of the two major error type among the 50 analysed incorrect term pairs

	Financial	Automotive	IT
Different meaning	38%	12%	**56%**
Partially correct	**62%**	**88%**	44%

5. Conclusions and future work

This paper describes TermEnsembler, a terminology extraction and alignment system, created from the point of view of language service providers in the language and translation industry. It consists of a concept-oriented terminology database with industry-standard file format support for easy sharing with other terminological applications, an online user interface for database management and semi-automatic term extraction, a monolingual terminology extraction algorithm (currently supporting English and Slovenian) and a novel bilingual alignment methodology with several components.

The first step is monolingual extraction based on the work of Vintar (2010) and Pollak et al. (2012) with some additional modifications, such as a filter for nested terms and near-duplicate recognition. The final result of this step are two lists of terms (one for each language) with the terms ordered by their termhood score. The next step, which is the central part of the paper, involves bilingual

alignment of the terms in the two lists. We have implemented and evaluated a total of seven methods – implementing approaches from the related work and the newly proposed approaches – which all return a list of aligned English-Slovenian term pairs. The evaluation of each approach separately shows that the highest precision was obtained by the newly developed phrase-table-based term alignment approach PTBA-3 which directly matches the extracted terms with phrases from the phrase table.

For final implementation, we experimented with different merging methods for the 7 outputs by assigning weights to produce a final list of term pairs. After initial experiments with equal weight and precision-based weights, we opted for an ensemble optimization approach using the genetic algorithm implementation from the evolutionary algorithm framework DEAP by Fortin et al. (2012), which takes random weight configurations and tries to optimize them towards a certain goal over a successive number of generations.

We have trained the bilingual alignment approach in TermEnsembler on one domain and tested it on two different domains achieving excellent results, with more than 96% of the top 400 term pair alignments produced by the system evaluated as correct by a human evaluator. In addition, we have also tried to optimize the system for producing a greater number of multi-word terms because they are particularly complicated for translation. When optimizing the evolutionary algorithm for overall precision and number of correct multi-word terms, at least a third of the top 400 term pair alignments returned by our system were correct multi-word terms, with precision computed on the MWUs reaching 0.919. All in all, we believe the high precision of our system among the top 400 terms would require only minor manual human curation to produce a viable term list for day-to-day work in the language industry.

We also briefly looked into whether bilingual term alignment improves the quality of monolingual terms. An experienced translator compared the top 200 terms returned by the initial algorithm (the LUIZ-CF variant described in Pollak et al. (2012)) for each of the two languages in all three domains and compared them with the top 200 terms produced by TermEnsembler after bilingual term alignment. The results show that TermEnsembler does improve the monolingual quality of terms (precision) by around 10%.

In terms of future work, we have identified several lines of research. We will continue adding new languages, implementing and systematically evaluating different monolingual term-extraction approaches. For bilingual alignment, we will initially focus on a systematic optimization of the evolutionary algorithm parameters and then look into implementing user-friendly parameters that would allow the users to tweak the weights towards greater overall precision or larger number of MWU terms. We will also test other, potentially faster optimization methods such

as differential evolution and Newton-like methods as well as develop machine-learning solutions for term alignment, combining the proposed statistical scores and cognate-based features, as in Aker et al. (2013). Finally, given a recent trend of well performing word-embeddings methods leading to excellent results in various natural-language processing tasks, we aim to address bilingual term-extraction as a well-suited task for developing cross-lingual embedding based term alignment methods, stimulated by the work of Conneau et al. (2018).

Acknowledgements

The system's interface and the elementary term extraction approaches were designed and developed in the scope of the TermIolar project by the Jožef Stefan Institute and Iolar d.o.o. The authors acknowledge the contribution of Simon Bratina and Davorin Sečnik (of Iolar d.o.o.) to functional specifications, additional requirements, evaluation of the interim results and providing important feedback and suggestions. The authors thank also Špela Vintar for her clarifications in the reimplementation of bilingual LUIZ term alignment.

The authors acknowledge the financial support of Slovenian Research agency for funding part of this research in the scope of basic research program Knowledge Technologies (Grant No. P2-0103) and the project Terminology and Knowledge Frames across Languages (Grant No. J6-9372). This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

References

Ahmad, Khurshid, Lee Gillam, and Lena Tostevin. 2000. "Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)." In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, 717–724. Washington, USA.

Aker, Ahmet, Monica Paramita, and Rob Gaizauskas. 2013. "Extracting Bilingual Terminologies from Comparable Corpora." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 402–411. Sofia, Bulgaria.

Amjadian, Ehsan, Diana Inkpen, Tahereh Paribakht, and Farahnaz Faez. 2016. "Local-Global Vectors to Improve Unigram Terminology Extraction." In *Proceedings of the 5th International Workshop on Computational Terminology*, 2–11. Osaka, Japan.

Baisa, Vít, Barbora Ulipová, and Michal Cukr. 2015. "Bilingual Terminology Extraction in Sketch Engine." In *9th Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2015 – Proceedings*, 61–67. Karlova Studánka, Czech Republic.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol: O'Reilly Media Inc.

Church, Kenneth Ward, and Patrick Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1): 22–29.

Cohen, Jacob. 1968. "Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin* 70 (4): 213. https://doi.org/10.1037/h0026256

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. "Word Translation Without Parallel Data." (https://arxiv.org/abs/1710.04087) Accessed 2 February 2019.

Daille, Béatrice, and Emmanuel Morin. 2005. "French-English Terminology Extraction from Comparable Corpora." In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, 707–718. Jeju Island, South Korea.

Daille, Béatrice, Éric Gaussier, and Jean-Marc Langé. 1994. "Towards Automatic Extraction of Monolingual and Bilingual Terminology." In *Proceedings of the 15th Conference on Computational linguistics*, 515–521. Kyoto, Japan. https://doi.org/10.3115/991886.991975

Dice, LR. 1945. "Measures of the Amount of Ecologic Association between Species." *Ecology* 26 (3): 297–302. https://doi.org/10.2307/1932409

Foo, Jody. 2012. *Computational Terminology: Exploring Bilingual and Monolingual Term Extraction*. Linköping: Linköping University Electronic Press.

Fortin, Félix-Antoine, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. "DEAP: Evolutionary Algorithms Made Easy." *Journal of Machine Learning Research* 13 (no. Jul): 2171–2175.

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mirna. 2000. "Automatic Recognition of Multi-Word Terms:. the C-Value/NC-Value Method." *International Journal on Digital Libraries* 3(2): 115–130. https://doi.org/10.1007/s007999900023

Gouadec, Daniel. 2007. *Translation as a Profession*. Amsterdam/Philadephia: John Benjamins. https://doi.org/10.1075/btl.73

Haque, Rejwanul, Sergio Penkale, and Andy Way. 2014. "Bilingual Termbank Creation via Log-Likelihood Comparison and Phrase-Based Statistical Machine Translation." In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, 42–51. Dublin, Ireland. https://doi.org/10.3115/v1/W14-4806

Hazem, Amir, and Emmanuel Morin. 2017. "Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora." In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, 685–693. Taipei, Taiwan.

Hiemstra, Djoerd. 1998. "Multilingual Domain Modeling in Twenty-One: Automatic Creation of a Bi-Directional Translation Lexicon from a Parallel Corpus." In *Proceedings of the 8th CLIN Meeting*, 41–58. Amsterdam, The Netherlands.

Justeson, John, and Slava Katz. 1995. "Technical Terminology: some Linguistic Properties and an Algorithm for Identification in Text." *Natural Language Engineering* 1 (1): 9–27. https://doi.org/10.1017/S1351324900000048

Kageura, Kyo, and Bin Umino. 1996. "Methods of Automatic Term Recognition: A Review." *Terminology* 3 (2): 259–289. https://doi.org/10.1075/term.3.2.03kag

Khan, Muhammad Tahir, Yukun Ma, and Jung-jae Kim. 2016. "Term Ranker: A Graph-Based Re-Ranking Approach." In *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference*, 310–315. Key Largo, USA.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan et al. 2007. "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 177–180. Prague, Czech Republic. https://doi.org/10.3115/1557769.1557821

Kupiec, Julian. 1993. "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora." In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, 17–22. Columbus, USA. https://doi.org/10.3115/981574.981577

Landis, Richard, and Gary Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–174. https://doi.org/10.2307/2529310

Ljubešić, Nikola, and Tomaž Erjavec. 2016. "Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene." In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 23–28. Portorož, Slovenia.

Logar, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, and Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba [Slovenian language corpora Gigafida, KRES, ccGigafida, ccKRES: creation, content, use]*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Macken, Lieve, Els Lefever, and Veronique Hoste. 2013. "Texsis: Bilingual Terminology Extraction from Parallel Corpora using Chunk-Based Alignment." *Terminology* 19 (1): 1–30. https://doi.org/10.1075/term.19.1.01mac

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Taylor & Francis.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." (https://arxiv.org/abs/1301.3781) Accessed 10 July 2018.

Neubig, Graham, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. "An Unsupervised Model for Joint Phrase Alignment and Extraction." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 632–641. Portland, USA.

Och, Franz Josef, and Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics* 29 (1): 19–51. https://doi.org/10.1162/089120103321337421

Pollak, Senja, Anže Vavpetič, Janez Kranjc, Nada Lavrač, and Špela Vintar. 2012. "NLP Workflow for On-Line Definition Extraction from English and Slovene Text Corpora." In *Proceedings of KONVENS 2012*, 53–60. Vienna, Austria.

Repar, Andraž, and Senja Pollak. 2017a. "Good Examples for Terminology Databases in Translation." In *Electronic Lexicography in the 21st century. Proceedings of eLex 2017 Conference*, 651–661. Leiden, Netherlands.

Repar, Andraž, and Senja Pollak. 2017b. "Ontology-Based Translation Memory Maintenance." In *Proceedings of the 20th International Multiconference Information Society 2017*, 19–22. Ljubljana, Slovenia.

Schmitz, Klaus Dirk, and Daniela Straub. 2016. "Tight Budgets and a Growing Number of Languages Impede Terminology Work." *tcworld magazine for international information management* (http://www.tcworld.info/e-magazine/technical-communication/article/tight-budgets-and-a-growing-number-of-languages-impede-terminology-work/). Accessed 24 August 2018.

The British National Corpus, version 3 (BNC XML Edition). 2007. *Distributed by Bodleian Libraries*, University of Oxford, on behalf of the BNC Consortium. (URL: http://www.natcorp.ox.ac.uk/). Accessed 10 March 2017.

Vintar, Špela. 2010. "Bilingual Term Recognition Revisited. The Bag-of-Equivalents Term Alignment Approach." *Terminology* 16 (2): 141–158. https://doi.org/10.1075/term.16.2.01vin

Wang, Rui, Wei Liu, and Chris McDonald. 2016. "Featureless Domain-Specific Term Extraction with Minimal Labelled Data." In *Proceedings of the Australasian Language Technology Association Workshop*, 103–112. Melbourne, Australia.

Wermter, Joachim, and Udo Hahn. 2005. "Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms." In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 843–850. Vancouver, Canada.

Wüster, Eugene. 1979. *Introduction to the General Theory of Terminology and Terminological Lexicography*. Vienna: Springer.

Zhang, Zigi, Jie Gao, and Fabio Ciravegna. 2018. "SemRe-Rank: Incorporating Semantic Relatedness to Improve Automatic Term Extraction Using Personalized PageRank." (https://arxiv.org/abs/1711.03373) Accessed 7 January 2019.

Address for correspondence

Andraž Repar
International Postgraduate School Jožef Stefan
Jožef Stefan Institute
Jamova 39, Ljubljana
Slovenia

repar.andraz@gmail.com

Co-author information

Vid Podpečan
Jožef Stefan Institute
vid.podpecan@ijs.si

Anže Vavpetič
Jožef Stefan Institute
hi@anzevavpetic.com

Nada Lavrač
Jožef Stefan Institute
nada.lavrac@ijs.si

Senja Pollak
Jožef Stefan Institute
senja.pollak@ijs.si

–
–

Chapter 3

Bilingual Terminology Alignment from Comparable Corpora

This chapter presents a comprehensive solution for bilingual terminology alignment from comparable corpora. After initially focusing on parallel corpora, which are the most common in the translation industry, we switched attention to comparable corpora, which do not have aligned segments, but are instead composed of texts from the same domain in two or more languages. As such, they do not contain direct translations, but are still a valuable source of bilingual (or multilingual) terminology.

In Section 3.1 we introduce the topic and define the problem as well as describe the reproducibility and replicability aspects of our research. Our initial approach to bilingual terminology alignment, starting from a simple reimplementation of an existing approach and then adding novel elements, is described in Section 3.2. In subsequent sections, we describe further modifications of our approach using new features or applying it to new domains. The relevant papers for this chapter are:

Repar, A., Martinc, M., & Pollak, S. (2020). Reproduction, replication, analysis and adaptation of a term alignment approach. Language Resources and Evaluation, 54(3), 767-800.
Code available here: http://source.ijs.si/mmartinc/4real2018.

Repar, A., Martinc, M., Ulčar, M., & Pollak, S. (2021). Word-embedding based bilingual terminology alignment. Proceedings of Electronic lexicography in the 21st century: Post-editing lexicography, 408-417.
Code available here: https://github.com/andrazrepar/terminology-alignment.

Repar, A., Pollak, S., Ulčar, M., & Koloski, B. (2022). Fusion of linguistic, neural and sentence-transformer features for improved term alignment. Proceedings of the BUCC Workshop within LREC 2022, 61-66.
Code available here: https://github.com/andrazrepar/terminology-alignment.

Repar, A., & Shumakov, A. (2021). Aligning Estonian and Russian news industry keywords with the help of subtitle translations and an environmental thesaurus. Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, 71–75.
Code available here: https://github.com/andrazrepar/terminology-alignment.

3.1 Introduction

Term alignment is most often one of the two steps of a single process: after first extracting terms using a (monolingual) term extraction approach in two languages, (bilingual) term alignment methods are used to align the two resulting term lists. The entire process can be called *bilingual terminology extraction*[1]. For the translation industry, bilingual term (extraction and) alignment from parallel corpora is extremely relevant due to the large amounts of sentence-aligned parallel corpora, but the translation memories containing these corpora are often owned by private companies who consider them their intellectual property and are reluctant to share them publicly. For this reason, considerable research efforts have also been invested into bilingual term (extraction and) alignment from comporable corpora, which are easier to compile.

Our work started with a reimplementation of an existing approach by Aker et al. (2013) using a machine learning methodology. During reimplementation, we initially encountered several areas where the description of the approach was not clear enough and we were unable to replicate the results. Hence, we expanded our research to include specific aspects of replicability in the field of terminology extraction and alignment, focusing on the availability of open datasets and code, as well as including an analysis of past papers from the field from the point of view of reproducibility and replicability.

We treat bilingual terminology alignment as a bilingual classification problem. Given two lists of terms, we create a dataset by pairing each term from List A with each term from List B, thereby generating AxB term pair candidates. We then create a training set and a held out test set, train a binary classifier and test it on the held out test set.

This section is divided into several subsections, each based on a paper using the above methodology. Section 3.2 describes the initial experiments using dictionary and cognate-based features, Section 3.3 describes experiments with additional word-embedding features, Section 3.4 describes experiments involving adding sentence-transformer features and Section 3.5 describes our attempts at applying the methodology to another domain.

3.2 Classification with Dictionary and Cognate-based Features

3.2.1 Description of the approach

The initial work, described in Repar et al. (2020), was based on an existing approach developed by Aker et al. (2013) to align terminology from comparable corpora using machine-learning techniques. They used terms from the EU's multilingual thesaurus Eurovoc (Steinberger et al., 2002) and train an SVM binary classifier (Joachims, 2002) The task of bilingual alignment is treated as a binary classification - each term from the source language S is paired with each term from the target language T and the classifier then decides whether the aligned pair is correct or incorrect. They then extract features (dictionary and cognate-based) to be used by the classifier. They run their experiments on the 21 official EU languages covered by Eurovoc with English always being the source language (20 language pairs altogether). They evaluate the performance on a held-out term pair list from Eurovoc using recall, precision and F_1 for all 20 languages. They report excellent performance on the held-out term list with many language pairs reaching 100% precision and the lowest recall being 65%. For Slovene, which was of our main interest, as well as for the additional target languages that we selected, namely French and Dutch, the reported

[1]Note that there is no clear consensus on naming and various papers used different names to refer to the same processes.

results were excellent with perfect or nearly perfect precision and good recall for all three language pairs.

The evaluation on the test set created as described in the original paper showed that compared to the results reported by the authors, our results were significantly worse. Despite all our efforts to follow the original approach, we were unable to match the results achieved in the original paper when running the algorithm without any changes to the original approach. For the EN-SL language pair, the reported results have the precision of 100% and the recall of 66%, meaning that with 600 positive term pairs in the test set, their classifier returns only around 400 positive term pairs. In contrast, in our replication attempts the classifier returned a lot of falsely classified positive term pairs. In addition to 526 true positive examples (out of a total of 600), the classifier also returned 14,194 misclassified examples - incorrect term pairs wrongly classified as correct. Similar statistics can be observed for the other two language pairs.

In order to improve the performance, we have performed experiments with regard to the following aspects:

- Giza++ terms only: we used only those terms that can be found in the Giza++ training corpora (i.e. the DGT translation memory)
- Giza++ cleaning: we removed low-quality alignments from the output of Giza++ tool
- Lemmatization: we lemmatized the input corpora and used lemmas for all subsequent phases of the methodology
- Changing the ratio of positive/negative examples in the training set: instead of using a balanced training set, we experimented with different ration of positive vs. negative examples
- Training set filtering: we removed positive examples that had the same characteristics as negative examples from the training set.
- Removing the Needleman-Wunsch Distance (NWD) feature: we removed NWD from the list of features, because it appeared to have the opposite effect of what we expected
- Term length filtering: we changed the prediction from positive to negative for term pairs that did not have the same number of words
- Adding novel cognate-based features: since the original implementation tended to overestimate the importance of Giza++ features, we added new cognate-based features

The main contribution of this paper are the novel cognate-based features, effectively utilizing the similarity of terms across languages. This allowed the algorithm to propose additional true positive term pairs not found using just the original set of features.

3.2.2 Results

Using the techniques described in Section 3.2.1, we were able to approach the results reported by the original paper on the Eurovoc dataset, with the best F_1 score on the EN-SL language pair achieved by a training set filtering configuration with a positive:negative ratio of 1:10. While the novel cognate approach did not yield the best results, it was nevertheless close to the best performing configurations and a manual evaluation has shown that it returned term pairs not found by other configurations.

Lang Resources & Evaluation
https://doi.org/10.1007/s10579-019-09477-1

ORIGINAL PAPER - REPLICABILITY & REPRODUCIBILITY

Reproduction, replication, analysis and adaptation of a term alignment approach

Andraž Repar[1,2] · **Matej Martinc**[1,2] · **Senja Pollak**[2,3]

Abstract In this paper, we look at the issue of reproducibility and replicability in bilingual terminology alignment (BTA). We propose a set of best practices for reproducibility and replicability of NLP papers and analyze several influential BTA papers from this perspective. Next, we present our attempts at replication and reproduction, where we focus on a bilingual terminology alignment approach described by Aker et al. (Extracting bilingual terminologies from comparable corpora. In: Proceedings of the 51st annual meeting of the association for computational linguistics, vol. 1 402–411, 2013) who treat bilingual term alignment as a binary classification problem and train an SVM classifier on various dictionary and cognate-based features. Despite closely following the original paper with only minor deviations—in areas where the original description is not clear enough—we obtained significantly worse results than the authors of the original paper. We then analyze the reasons for the discrepancy and describe our attempts at adaptation of the approach to improve the results. Only after several adaptations, we achieve results which are close to the results published in the original paper. Finally, we perform the experiments to verify the replicability and reproducibility of our own code. We publish our code and datasets online to assure the reproducibility of the results of our experiments and implement the selected BTA models in an online

✉ Andraž Repar
repar.andraz@gmail.com

Matej Martinc
matej.martinc@ijs.si

Senja Pollak
senja.pollak@ijs.si

[1] Jožef Stefan Postgraduate School, Ljubljana, Slovenia

[2] Jožef Stefan Institute, Ljubljana, Slovenia

[3] Usher institute, Medical school, University of Edinburgh, Edinburgh, UK

Published online: 18 November 2019

platform making them easily reusable even by the technically less-skilled researchers.

Keywords Bilingual term alignment · Reproducibility · Machine learning · Cognates

1 Introduction

The issue of reproducibility has been on the radar of researchers at least for the past 25 years, particularly in the life science research (e.g. Yentis et al. 1993; Prinz et al. 2011; Camerer et al. 2016). More recently, many other disciplines have started to acknowledge the crisis of reproducibility, among them also human language technology research (Pedersen 2008; Kano et al. 2009; Fokkens et al. 2013; Branco et al. 2017; Wieling et al. 2018). However, the basic terminology has remained confusing with different authors using different terms for the same concepts which is why Cohen et al. (2018) describe the three dimensions of reproducibility in natural language processing (NLP) and provide a set of definitions for the various concepts used when discussing reproducibility in NLP. They first differentiate between the concepts of **replicability** (or repeatability), which they define as *the ability to repeat the experiment described in a study*, and **reproducibility**, which describes the outcome—whether *the replicability efforts lead to the same conclusions.* Then they further break down reproducibility into reproducibility of a **conclusion** (defined as an explicit statement in the paper arrived at on the basis of the results of the experiments), reproducibility of a **finding** (a relationship between the values for some reported figure of merit) and reproducibility of a **value** (actual measured or calculated numbers).

In this paper we extend our reproducibility study (Repar et al. 2018), presented at the Workshop on Research Results Reproducibility and Resources Citation (4REAL Workshop, Branco et al. (2018)) organized within the scope of the 11th Language Resources and Evaluation Conference (LREC 2018). Our original motivation came from our interest and need for a terminology alignment tool, and the paper by Aker et al. (2013) titled "Extracting Bilingual Terminology from Parallel Corpora" seemed a perfect candidate for reproduction with nearly perfect results, coverage of the Slovenian-English pair (which were the languages of our interest) and what seemed like a well described and simple to replicate method. The authors treat aligning terms in two languages as a binary classification problem. They use an SVM binary classifier (Joachims 2002) and training data terms taken from the Eurovoc thesaurus (Steinberger et al. 2002) and construct two types of features: dictionary-based (using word alignment dictionaries created with Giza++ (Och and Ney 2003)) and cognate-based (effectively utilizing the similarity of terms across languages). Given that the results looked very promising—precision on the held-out set was 1 or close to 1 for many language pairs, we thought we could use the approach in our work and we set out to replicate it. We expected a straightforward process, but it turned out to be anything but: the results of our experiments were very vastly different from the original paper. For example, while the original paper

reports an extremely high precision (1 or close to 1) for the language pairs we have focused on, our experiments showed a precision below 0.05. Based on the reproducibility dimensions mentioned above, in our original reproducibility experiment from Repar et al. (2018) we were not able to reproduce any of the three dimensions: the values and findings in our experiments were vastly different, and—had we stopped at this point—we would have concluded that the proposed machine learning approach is not suitable for bilingual terminology alignment. Only after a great deal of tweaking and optimization have we managed to get to a respectable precision level (similar to the results in the original paper).

In the present paper, we aim to explore the issue of reproducibility and replicability in the field of terminology alignment further. To do so, we extend the work in Repar et al. (2018) with the following:

- an overview of bilingual terminology extraction and alignment approaches in terms of replicability and reproducibility.
- extending the original reproducibility experiment to two additional languages, resulting in Slovenian, French and Dutch as target languages from three different language families.
- providing very detailed description of feature construction.
- additional filtering and refinement of the cognate-based features.
- a reproducibility experiment with source code from Repar et al. (2018).
- implementation of our code into an online data mining platform ClowdFlows.
- a discussion on good practices for reproducibility and replicability in NLP.

This paper is organized as follows: After the introduction in Sect. 1, we present the related work and the analysis of bilingual terminology alignment papers from the point of view of replicability and reproducibility (Sect. 2). Section 3 contains the main replicability and reproducibility experiments, and is followed by Sect. 4, which describes our attempts at improving the results of the replicated approach, while Sect. 5 contains the results of manual evaluation. Section 6 describes the reproducibility experiment using our code from Repar et al. (2018) and Sect. 7 the implementation of the system in the ClowdFlows platform, for making it accessible to a wider community. Section 8 contains the conclusions and presents ideas for future work. The code and datasets of our experiments are published online, to enable future reproducibility and replicability.[1]

2 Overview of bilingual terminology extraction and alignment approaches

In this section we first look at the related work on bilingual terminology extraction and alignment and then analyze several related papers from the viewpoint of replicability and reproducibility.

[1] http://source.ijs.si/mmartinc/4real2018.

2.1 Related work

We start by providing a clarification regarding the terminology used in this paper. Following the distinction between two basic approaches made by Foo (2012):

- *extract-align* where we first extract monolingual candidate terms from both sides of the corpus and then align the terms, and
- *align-extract* where we first align single and multi-word units in parallel sentences and then extract the relevant terminology from a list of candidate term pairs.

we propose the following two definitions:

- *Bilingual terminology extraction* is the process which, given the input of related specialized monolingual corpora, results in the output of terms aligned between two languages. The process can either start with extracting monolingual candidate terms and aligning them between two languages (i.e. extract-align) or with aligning phrases and then extracting terms (i.e. align-extract) or any other sequence of actions.
- *Bilingual terminology alignment* is the process of aligning terms between two candidate term lists in two languages.

Bilingual terminology alignment has a narrower focus than bilingual terminology extraction, but the two terms are often used interchangeably in various papers. For example, the title of the paper we were trying to replicate "Extracting bilingual terminologies from comparable corpora" is somewhat misleading in this regard, since the paper primarily deals with bilingual terminology alignment, while they utilize monolingual terminology extraction (specifically the approach by Pinnis et al. (2012) without any modifications) only in the manual evaluation experiments.

The primary purpose of bilingual terminology extraction is to build a term bank—i.e. a list of terms in one language along with their equivalents in the other language. With regard to the input text, we can distinguish between alignment on the basis of a parallel corpus and alignment on the basis of a comparable corpus. For the translation industry, bilingual terminology extraction from parallel corpora is extremely relevant due to the large amounts of sentence-aligned parallel corpora available in the form of translation memories (in the TMX file format). Consequently, initial attempts at bilingual terminology extraction involved parallel input data (Kupiec 1993; Daille et al. 1994; Gaussier 1998), and the interest of the community continued until today (Ha et al. 2008; Ideue et al. 2011; Macken et al. 2013; Haque et al. 2014; Arčan et al. 2014; Baisa et al. 2015). However, most parallel corpora are owned by private companies,[2] such as language service providers, who consider them to be their intellectual property and are reluctant to share them publicly. For this reason (and in particular for language pairs not

[2] However, some publicly available parallel corpora do exist. A good overview can be found at the OPUS web portal (Tiedemann 2012).

involving English) considerable efforts have also been invested into researching bilingual terminology extraction from comparable corpora (Fung and Yee 1998; Rapp 1999; Chiao and Zweigenbaum 2002; Cao and Li 2002; Daille and Morin 2005; Morin et al. 2008; Vintar 2010; Bouamor et al. 2013; Hazem and Morin 2016, 2017).

Despite the problem of bilingual term alignment lending itself well to the binary classification task, there have been relatively few approaches utilizing machine learning. For example, similar to Aker et al. (2013), Baldwin and Tanaka (2004) generate corpus-based, dictionary-based and translation-based features and train an SVM classifier to rank the translation candidates. Note that they only focus on multi-word noun phrases (noun + noun). A similar approach, again focusing on noun phrases, is also described by Cao and Li (2002). Finally, Nassirudin and Purwarianti (2015) also reimplement Aker et al. (2013) for the Indonesian-Japanese language pair and further expand it with additional statistical features. In the best scenario, their accuracy, precision and recall all exceed 90% but the results are not directly comparable since Nassirudin and Purwarianti (2015) use tenfold cross-validation while Aker et al. (2013) use a held-out test set. In addition, Nassirudin and Purwarianti (2015) have a balanced test set while Aker et al. (2013) use a very unbalanced one (ratio of positive vs. negative examples 1:2000).

2.2 Analysis of past papers on bilingual terminology extraction from the viewpoint of reproducibility and replicability

In an ideal reproducibility and replicability scenario, a scientific paper would contain an accurate and clear description of the datasets used and experiments conducted and the authors would provide a single link containing all the datasets (versions, subsets etc.) used for the experiments along with the experiment source code (or alternatively, an online tool to run the experiments). These could then be used to replicate the experiments and reproduce the results using the descriptions provided in the paper.

We have analyzed several[3] bilingual terminology extraction papers from the past 25 years from the point of view of dataset, code and tool availability. The summary of results is available in Table 1.

2.2.1 Dataset availability

In terms of dataset availability, we looked at whether the paper contains some description of how the datasets were constructed and which could (theoretically) be used to reconstruct the datasets. Note that under "dataset", we include corpora, gold

[3] The selection process was as follows: the starting point were selected seminal papers on the field, as well as two queries in the ACL Anthology database: "term alignment" and "bilingual terminology extraction". We analyzed the papers found by these two queries as well as additional papers mentioned in the related works sections of these papers and the main criterion for including a paper in our analysis was that it primarily deals with bilingual terminology extraction (and not for example latent semantic analysis, such as Bader and Chew (2008)). However, no strict systematic review with inclusion and exclusion criteria was made, as such a survey would be beyond the needs of this paper.

Table 1 An analysis of bilingual terminology extraction papers from the point of view of reproducibility and replicability

Paper	Dataset	Code	Tool	Google Scholar citations as of September 2019
Kupiec (1993)	Links	No	No	333
Daille et al. (1994)	No	No	No	268
Fung and Yee (1998)	Description	No	No	427
Gaussier (1998)	No	No	No	84
Rapp (1999)	Description	No	No	552
Chiao and Zweigenbaum (2002)	Description	No	No	135
Cao and Li (2002)	Description	No	No	141
Morin et al. (2007)	No	No	No	113
Daille and Morin (2005)	Obsolete	No	Obsolete	56
Morin et al. (2008)	Links	No	Obsolete	22
Ha et al. (2008)	Description	No	No	4
Lee et al. (2010)	Description	No	No	22
Vintar (2010)	No	No	Obsolete	53
Ideue et al. (2011)	No	No	Yes[a]	9
Macken et al. (2013)	No	No	No	48
Bouamor et al. (2013)	Description	No	No	24
Aker et al. (2013)	Links	No	No	36
Arčan et al. (2014)	Links	No	No	18
Haque et al. (2014)	Links	No	No	11
Kontonatsios et al. (2014)	Description	No	No	14
Baisa et al. (2015)	No	No	Yes	5
Hazem and Morin (2016)	Links	No	No	12
Hazem and Morin (2017)	Links	No	No	2

[a]A Perl module (Term Extract) was used, however the link leads to a Japanese website

standard termlists, seed dictionaries and all other linguistic resources needed to conduct the experiments in the paper. For example, we consider the following paragraph from Rapp (1999) to be a valid description of a dataset: *As the German corpus, we used 135 million words of the newspaper Frankfurter Allgemeine Zeitung (1993 to 1996), and as the English corpus 163 million words of the Guardian (1990 to 1994).* On the other hand, this paragraph from Ideue et al. (2011) is not considered a valid description: *We extracted bilingual term candidates from a Japanese-English parallel corpus consisting of documents related to apparel products.* In the former example, dataset reconstruction would be difficult but not impossible, while in the latter it is impossible. An even better option is to link to actual datasets or refer to papers where datasets are described and linked, which is why we also looked for dataset links and/or references in the analyzed papers. Note that there are several examples where links are provided only for a selection of the datasets used in the experiments (e.g., Morin et al. (2008)).

As evident from Table 1, dataset availability is the least problematic aspect of reproducibility and replicability in terminology (extraction and) alignment papers with approximately two thirds of the analyzed papers (15 out of 23) either containing a description of the resources used for the experiments, providing links to them or refering to papers where they are described.

We expected the earlier papers to have less information on datasets than latter ones, but this turned out not to be the case. In fact, the earliest paper analyzed—Kupiec (1993)—provides a reference to a publicly available corpus (Canadian Hansards (Gale and Church 1993)). The first paper to have a separate section with data/resource description is Rapp (1999) and from this point on, almost all papers have such a section—usually titled "Data and Resources", "Resources and Experimental Setup", "Linguistic resources" or similar.

However, it is rarely documented what version of the dataset was used and whether an entire dataset was used or only a part of it (as in random selection, train-test split, etc.). In most cases, little information is provided on the actual subsets used for the experiments. Another aspect of dataset use is the languages: when one of the languages involved is English, it is much easier to find datasets than for other language combinations. Finally, there is also the issue of keeping the links active. For example, many of the links in Daille and Morin (2005) and Morin et al. (2008) are not active anymore while Bouamor et al. (2013) state that the corpora and terminology gold standard lists created for the paper will be shared publicly, but no links are provided.

The most significant problem encountered during our analysis was the fact that terminology alignment is most often not the sole focus of a paper, such as in Haque et al. (2014), where the experiments start with monolingual terminology extraction from two languages and the extracted terms are then aligned. As terminology extraction and alignment go hand-in-hand, it may often be impossible to make a clear distinction between the terminology extraction and terminology alignment datasets. This means that the dataset results in Table 1 are not a true apple-to-apple comparison: one paper might link to the parallel corpus used to extract terms from, while another to a gold standard termlist. Our main criterion was whether the dataset description (or link) could be used to replicate the experiments described in the paper.

An ideal terminology (extraction and) alignment dataset would therefore consist of a bilingual or multilingual (parallel or comparable) corpus along with reference (gold standard) term lists containing terms that can be found in the corpus. Such corpora are TTC wind energy and TC mobile technology[4], which contain data for six languages (English, French, German, Spanish, Russian, Latvian, Chinese), or the Bitter corpus[5], which contains data for the EN-IT language pair. The first was used in Hazem and Morin (2016), while the second one by Arčan et al. (2014). Since such datasets are scarce, researchers employ various methodologies for constructing their own datasets. One method, used by Aker et al. (2013), is to take one of the available multilingual translation memories containing EU documentation (such as

[4] http://www.lina.univ-nantes.fr/?Reference-Term-Lists-of-TTC.html.

[5] https://hlt-mt.fbk.eu/technologies/bittercorpus.

Europarl (Koehn 2005) or DGT (Steinberger et al. 2013)) as the corpus and a glossary (e.g., IATE (Johnson and Macphail 2000)) or thesaurus (e.g., Eurovoc (Steinberger et al. 2002)) as the terminology gold standard list. Another strategy, used by Hazem and Morin (2017), is to collect a comparable corpus manually (i.e. scientific articles in French and English from the Elsevier[6] website) and a domain specific terminological resource (i.e. UMLS[7]) as a reference termlist. Hazem and Morin (2017) also filter out those terms from the termlist that do not appear often enough in their corpus. In other cases (e.g., Haque et al. (2014)), the datasets are not available because the papers were written as part of industrial projects and the datasets are private.

2.2.2 Code and tool availability

We have discovered that no paper has made experiment code available and only a few provide access or links to tools where the experiments were conducted. But even when links to tools are provided, reproducibility and replicability may be hindered: for example, the link provided in Ideue et al. (2011) leads to a Japanese website. Another issue is the long-term availability of resources. For example, Daille and Morin (2005) conducted their experiments in *ACABIT*, an open source terminology extraction software. However, the link given in the paper does not work anymore. From the analyzed papers, the only example of bilingual term extraction and alignment tool, which is publicly available, is the Sketch Engine term extraction module, described by Baisa et al. (2015).

None of the papers analyzed in this section fulfill the ideal scenario described at the start of this section (i.e. a single link with code and all datasets) which severly hinders any replicability attempts as will be evident from our own experiments described in this paper.

3 Replicating a machine learning approach to bilingual term alignment and reproducing its results

This section describes our efforts in replicating a machine learning approach to bilingual term alignment described in Aker et al. (2013),by which we extend our initial experiments and analysis (Repar et al. 2018). Section 3.1 describes the original approach and Sect. 3.2 contains an overview of our attempts to replicate it.

3.1 Description of the original approach

The original approach designed by Aker et al. (2013) was developed to align terminology from comparable (or parallel) corpora using machine-learning techniques. They use terms from the Eurovoc (Steinberger et al. 2002) thesaurus and train an SVM binary classifier (Joachims 2002) (with a linear kernel and the

[6] https://www.elsevier.com/.

[7] https://www.nlm.nih.gov/research/umls/.

trade-off between training error and margin parameter c = 10). The task of bilingual alignment is treated as a binary classification—each term from the source language *S* is paired with each term from the target language *T* and the classifier then decides whether the aligned pair is correct or incorrect. They then extract features (dictionary and cognate-based) to be used by the classifier. They run their experiments on the 21 official EU languages covered by Eurovoc with English always being the source language (20 language pairs altogether). They evaluate the performance on a held-out term pair list from Eurovoc using recall, precision and F-measure for all 20 languages. Next, they propose an experimental setting for a simulation of a real-world scenario where they collect English-German comparable corpora of two domains (IT, automotive) from Wikipedia, perform monolingual term extraction using the system by Pinnis et al. (2012) followed by the bilingual alignment procedure described above and manually evaluate the results (using two evaluators). They report excellent performance on the held-out term list with many language pairs reaching 100% precision and the lowest recall being 65%. For Slovenian, which is of our main interest, as well as for the additional target languages that we selected, namely French and Dutch, the reported results were excellent with perfect or nearly perfect precision and good recall for all three language pairs. The reported results of the manual evaluation phase were also good, with two evaluators agreeing that at least 81% of the extracted term pairs in the IT domain and at least 60% of the extracted term pairs in the automotive domain can be considered exact translations.

3.1.1 Features

Aker et al. (2013) use two types of features that express correspondences between the words (composing a term) in the target and source language (for a detailed description see Table 2:

– 7 dictionary-based (using Giza++) features which take advantage of dictionaries created from large parallel corpora of which 6 are direction-dependent (source-to-target or target-to-source) and 1 direction-independent—resulting in altogether 13 features, and

– 5 cognate-based (on the basis of Gaizauskas et al. (2012)) which utilize string-based word similarity between languages.

To match words with morphological differences, they do not perform direct string matching but utilize Levenshtein Distance. Two words were considered equal if the Levenshtein Distance (Levenshtein 1966) was equal or higher than 0.95. For closed-compounding languages, they check whether the compound source term has an initial prefix that matches the translation of the first target word, provided that translation is at least 5 characters long.

Table 2 Features used in the experiments

Feature	Cat	Description	Type
isFirstWordTranslated	Dict	Checks whether the first word of the source term is a translation of the first word in the target term (based on the Giza++ dictionary)	Bin
isLastWordTranslated	Dict	Checks whether the last word of the source term is a translation of the last word in the target term	Bin
percentageOfTranslatedWords	Dict	Ratio of source words that have a translation in the target term	Num
percentageOfNotTranslatedWords	Dict	Ratio of source words that do not have a translation in the target term	Num
longestTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which has a translation in the target term (compared to the source term length)	Num
longestNotTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which do not have a translation in the target term (compared to the source term length)	Num
Longest Common Subsequence Ratio (LCSSR)	Cogn	Measures the longest common non-consecutive sequence of characters between two strings (divided by the length of the longest string)	Num
Longest Common Substring Ratio (LCSTR)	Cogn	Measures the longest common consecutive string (LCST) of characters that two strings have in common (divided by the length of the longest string)	Num
Dice similarity	Cogn	2*LCST / (len(source) + len(target))	Num
Needlemann-Wunsch distance	Cogn	LCST / min(len(source), len(target))	Num
Normalized Levenshtein distance (nLD)	Cogn	1 − LD / max(len(source), len(target))	Num
isFirstWordCovered	Comb	A binary feature indicating whether the first word in the source term has a translation or transliteration in the target term	Bin
isLastWordCovered	Comb	A binary feature indicating whether the last word in the source term has a translation or transliteration in the target term	Bin
percentageOfCoverage	Comb	Returns the percentage of source term words which have a translation or transliteration in the target term	Num
percentageOfNonCoverage	Comb	Returns the percentage of source term words which have neither a translation nor transliteration in the target term	Num
difBetweenCoverageAndNonCoverage	Comb	Returns the difference between the last two features	Num

Note that some features are used more than once because they are direction-dependent

Additional features are also constructed by:

- Using language pair specific transliteration rules to create additional cognate-based features. The purpose of this task was to try to match the cognate terms while taking into account the differences in writing systems between two languages: e.g. Greek and English. Transliteration rules were created for both directions (source-to-target and target-to-source) separately and cognate-based features were constructed for both directions - resulting in additional 10 cognate-based features with transliteration rules.
- Combining the dictionary and cognate-based features in a set of combined features where the term pair alignment is correct if either the dictionary or the cognate-based method returns a positive result. This process resulted in additional 10 combined features.[8]

At the end of the feature construction phase, there were 38 features: 13 dictionary-based, 5 cognate-based, 10 cognate-based features with transliteration rules and 10 combined features.

3.1.2 Data source and experiments

Using Giza++, Aker et al. (2013) create source-to-target and target-to-source word alignment dictionaries based on the DGT translation memory (Steinberger et al. 2013). The resulting dictionary entries consist of the source word *s*, its translation *t* and the number indicating the probability that *t* is an actual translation of *s*. To improve the performance of the dictionary-based features, the following entries were removed from the dictionaries:

- entries where probability is lower then 0.05.
- entries where the source word was less than 4 characters and the target word more than 5 characters long and vice versa in order to avoid translations of stop word to content words.)

The next step is the creation of term pairs from the Eurovoc (Steinberger et al. 2002) thesaurus, which at the time consisted of 6797 terms. Each non-English language was paired with English. The test set consisted of 600 positive (correct) term pairs—taken randomly out of the total 6797 Eurovoc term pairs—and around 1.3 million negative pairs which were created by pairing each source term with 200 distinct incorrect random target terms. Aker et al. (2013) argue that this was done to simulate real-world conditions where the classifier would be faced with a larger number of negative pairs and a comparably small number of positive ones. The 600 positive term pairs were further divided into 200 pairs where both (i.e. source and target) terms were single words, 200 pairs with a single word only on one side and

[8] For combined features, a word is considered as covered if it can be found in the corresponding set of Giza++ translations or if one of the cognate-based measures (Longest Common Subsequence, Longest Common Substring, Levenshtein Distance, Needleman-Wunsch Distance, Dice) is 0.70 or higher (set experimentally by Aker et al. (2013)).

200 pairs with multiple-word terms on both sides. The remaining positive term pairs (approximately 6200) were used as training data along with additional 6200 negative pairs. These were constructed by taking the source side terms and pairing each source term with one target term (other than the correct one). Using this approach, Aker et al. (2013) achieve excellent results with 100% precision and 66% recall for Slovenian and French and 98% precision and 82% recall for Dutch.

3.2 Replication of the approach

The first step in our approach was to replicate the algorithm described by Aker et al. (2013). The initial premise is the same: given two lists of terms from the same domain in two different languages, we would like to align the terms in the two lists to get one bilingual glossary to be used in a variety of settings (computer-assisted translation, machine translation, ontology creation etc.). We followed the approach described above faithfully except in the following aspects[9]:

- Instead of the entire set of Eurovoc languages, we have initially focused only on the English-Slovenian language pair (Repar et al. 2018). In the current paper, we add two additional language pairs (English-French, English-Dutch) to see whether our findings can be generalised across different languages. We selected languages from different language families, as the importance of cognates is dependent on the similarity between languages (for example, Dutch and English (being both Germanic languages) presumably have a higher number of cognates).
- We use newer datasets. The Eurovoc thesaurus version that we used contained 7,083 terms for Slovenian[10] and 7,181 terms for French[11] and Dutch.[12] Similarly, the DGT translation memory contains additional content not yet present in 2013.[13] For English-Slovenian, we at first used the entire DGT corpus up to and including the *DGT-TM-release 2017* for deriving GIZA alignments. Later we also experimented with precomputed dictionaries by Aker et al. (2014). When performing the experiments on the other languages pairs, we did not create our own GIZA alignment, but only used the precomputed ones by Aker et al. (2014).
- Since no particular cleaning of training data (e.g., manual removal of specific entries) is described in the paper for the languages of our interest, we do not perform any.

We think that regardless of these differences, the experiments should yield similar results.

[9] Note that our original replication paper Repar et al. (2018) wrongly states that we did not utilize the compounding solution implemented by Aker et al. (2013) for addressing compouding issues in languages such as German. In fact, we did implement it and used it in all experiments.

[10] http://source.ijs.si/mmartinc/4real2018/blob/master/term_list_sl.csv.

[11] http://source.ijs.si/mmartinc/4real2018/blob/master/term_list_fr.csv.

[12] http://source.ijs.si/mmartinc/4real2018/blob/master/term_list_nl.csv.

[13] The versions of the resources used in Aker et al. (2013) were not documented or made available.

Reproduction, replication, analysis and adaptation...

3.2.1 Problems with replicating the approach

While the general approach is clearly laid out in the article, there are several spots where further clarification would be welcome:

- There is no sufficient information about the Giza++ settings or whether the input corpora have been lemmatized. In order to improve term matching, we experimented with and without lemmatization of the Giza++ input corpora.
- There is no information about the specific character mappings rules other than a general principle of one character in the source being mapped to one or more character in the target. Since the authors cover 20 languages, it is understandable that they cannot include the actual mapping rules in the article. Therefore, we have created our own mapping rules for English-Slovenian and English-French according to the instructions in the original paper:
  - Mapping the English term to the Slovenian writing system (the character before the colon is replaced by the sequence of characters after the colon): *x:ks, y:j, w:v, q:k.*
  - Mapping the Slovenian term to the English writing system: *č:ch, š:sh, ž:zh.*
  - Mapping the French term to the English writing system: we deleted all accents e.g., é:e, ê:e.
  - Mapping the Dutch term to the English writing system: we deleted all accents and replace the digraph ij with two separate letters ij.

- Instead of the unclear Needleman–Wunsch distance formula from Aker et al. (2013) $\frac{LCST}{min[len(source)+len(target)]}$ (which implies that we should take the minimum value of the sum of the length of the target and source term) we opted for $\frac{LCST}{min[len(source),len(target)]}$ as in Nassirudin and Purwarianti (2015).
- We were not completely certain how to treat examples such as "passport—potni list", where a single-word source term is translated by a multi-word target term and both combinations (passport—potni and passport—list) can be found in the Giza++ dictionary. In this case, our implementation returns values of 1 for both *isFirstWordTranslated* and *isLastWordTranslated* features despite the fact that the source term only has one word.
- There was a slight ambiguity on how to calculate cognate-based features: on the level of words or on the level of entire terms. We opted for the second, since the names of the cognate-based features did not imply that cognates are calculated on the word level (as was the case with the dictionary-based features) and since there was no mention in the original paper on how to combine cognate-based scores for specific word pairs in the multi-word term pairs in order to get a final cognate score for the whole term pair.
- In the original article, the *isFirstWordCovered* feature is described as "a binary feature indicating whether the first word in the source term has a translation (i.e. has a translation entry in the dictionary regardless of the score) or transliteration (i.e. if one of the cognate metric scores is above 0.7) in the target term." While

the dictionary-based part is clear, for calculating the cognate-based feature values (e.g., of the first word in the source term), the values of the cognate metric scores concern the entire target term. As we did not find this fully intuitive, and we believe other interpretations are possible, we experimented with these settings in the adaptation of the approach (see Sect. 4.8).
To avoid ambiguities, we provide a separate document with examples of constructed features, together with the code (http://source.ijs.si/mmartinc/4real2018/blob/master/feature_examples.docx).

3.2.2 Results

The evaluation on the test set created as described in the original paper by Aker et al. (2013) shows that compared to the results reported by the authors (see line 1 in Tables 3, 4 and 5), our results are significantly worse. Despite all our efforts to follow the original approach, we were unable to match the results achieved in the original paper when running the algorithm without any changes to the original approach. When trying to follow the original paper's methodology, precision is only 3.59% and recall is 88% for the English-Slovenian language pair. The results for the other two language pairs are comparable (see line 2 in Tables 3, 4 Table 5 for details).

In Sect. 4, we provide the results of detailed analysis and additional experiments that we performed in order to reach results comparable to the original approach.

3.2.3 Attempts at establishing contact with the authors

When replicating an existing paper, especially when the code is not made available, contacting the authors for clarification (or for providing/running the code) is the most obvious step when encountering the problems or ambiguities. However, due to busy schedules of researchers, change of professional paths or other similar reasons, getting detailed help might be impossible.

This is true for our case as well. Initially, we were hopeful of getting useful feedback, as the authors already provided the software to other researchers in the past (see Arčan et al. (2014)). However, despite a friendly response, we have been able to get only a limited number of answers and many questions remained unanswered, and the auhors have not been able to share their code. We have first contacted the original authors of the paper when we were running the experiments reported in Repar et al. (2018) and did receive some answers confirming our assumptions (e.g. regarding mapping terms to the different writing systems and that the test set data was selected individually for each language pair), but several other issues remained unaddressed (in particular, what was the exact train and test data selection strategy for the EN-SL language pair). Further inquiries proved unsuccessful due to time constraints on the part of the original authors. As we expanded the paper with additional languages and experiments, we again contacted the main author, provided him the code and the paper and asked for help in

Table 3 Results on the English–Slovenian term pair

No.	Config EN-SL	Training set size	Pos/neg ratio	Precision	Recall	F-score
1	Reported by Aker et al. (2013)	12,400	1:1	1	0.6600	0.7900
2	Replicated approach	12,966	1:1	0.0359	0.8800	0.0689
3	Giza++ terms only	8306	1:1	0.0645	**0.9150**	0.1205
4	Giza++ cleaning	12,966	1:1	0.0384	0.7789	0.0731
4a	Lemmatization	12,966	1:1	0.0373	0.8150	0.0713
5	Training set 1:200	1,303,083	1:200	0.4299	0.7617	0.5496
6	Training set filtering 1	6426	1:1	0.5969	0.64167	0.6185
7	Training set filtering 2	35,343	1:10	0.9042	0.5350	**0.6723**
8	Training set filtering 3	645,813	1:200	**0.9342**	0.4966	0.6485
9	Term length filtering	6426	1:1	0.8144	0.4900	0.6119
10	Cognates approach	672,345	1:200	0.8732	0.5167	0.6492

No. 1 presents the results reported by the authors, No. 2 our replication of the approach and No. 3–10 our modifications of the first replicated approach with the aim of improving the results

Table 4 Results on the English–French language pair

No.	Config EN-FR	Training set size	Pos/neg ratio	Precision	Recall	F-score
1	Reported by Aker et al. (2013)	12,400	1:1	1	0.6600	0.7900
2	Replicated approach	13,160	1:1	0.0323	**0.8483**	0.0622
3	Giza++ terms only	8892	1:1	0.0437	0.8433	0.0830
4	Giza++ cleaning	13,160	1:1	0.0317	0.7917	0.0610
5	Training set 1:200	1,322,580	1:200	0.5273	0.6767	**0.5927**
6	Training set filtering 1	2650	1:1	0.4623	0.5517	0.5030
7	Training set filtering 2	14,575	1:10	0.9422	0.3533	0.5139
8	Training set filtering 3	266,325	1:200	**0.9791**	0.3117	0.4728
9	Term length filtering	2650	1:1	0.6791	0.3950	0.4995
10	Cognates approach	311,952	1:200	0.8603	0.3900	0.5367

No. 1 presents the results reported by the authors, No. 2 our replication of the approach and No. 3–10 our modifications of the first replicated approach with the aim of improving the results

identification of any possible mistakes leading to the results, however, we were ultimately not able to get any information which would explain the differences.

We think the original paper is generally well-written and that the main reason for occasional lack of clarity is its scope: as the authors deal with more than 20 language pairs, it would be impossible to provide specific information regarding all of them. Providing more examples would be useful, but still the code and the exact dataset are in our opinion the only way to be able to fully replicate the experiments.

Table 5 Results on the English–Dutch language pair

No.	Config EN-NL	Training set size	Pos/neg ratio	Precision	Recall	F-score
1	Reported by Aker et al. (2013)	12,400	1:1	0.9800	0.8200	0.8000
2	Replicated approach	13,160	1:1	0.0227	0.8850	0.0442
3	Giza++ terms only	7310	1:1	0.0636	**0.9317**	0.1191
4	Giza++ cleaning	13,160	1:1	0.0340	0.8500	0.0654
5	Training set 1:200	1,322,580	1:200	0.5053	0.6300	0.5608
6	Training set filtering 1	4250	1:1	0.5122	0.4917	0.5017
7	Training set filtering 2	23,375	1:10	0.6842	0.4333	0.5306
8	Training set filtering 3	427,125	1:200	**0.9356**	0.3633	0.5234
9	Term length filtering	4250	1:1	0.7621	0.3683	0.4966
10	Cognates approach	468,933	1:200	0.9101	0.5233	**0.6646**

No. 1 presents the results reported by the authors, No. 2 our replication of the approach and No. 3–10 our modifications of the first replicated approach with the aim of improving the results

4 Analysis and adaptation: experiments for improving the replicated approach

The results in our replicated experiments differ dramatically from the results obtained by Aker et al. (2013). Their approach yields excellent results with perfect or almost perfect precision and respectable recall for all three languages under our consideration.

For the EN-SL language pair, the reported results have the precision of 100% and the recall of 66%, meaning that with 600 positive term pairs in the test set, their classifier returns only around 400 positive term pairs. In contrast, in our replication attempts the classifier returned a lot of falsely classified positive term pairs. In addition to 526 true positive examples (out of a total of 600), the classifier also returns 14,194 misclassified examples—incorrect term pairs wrongly classified as correct. Similar statistics can be observed for the other two language pairs.

These results are clearly not useful for our goals which is to use the methods to continuously populate a termbase with as little manual intervention as possible. In this section we present the analysis of ambiguities in the description of the approach and the issues spotted when inspecting the results of the replicated approach, and propose several methods aiming at improving the results. To do so, we have performed experiments with regard to the following aspects:

- Giza++ terms only: using only those terms that can be found in the Giza++ training corpora (i.e. DGT).
- Giza++ cleaning.
- Lemmatization.
- Changing the ratio of positive/negative examples in the training set.
- Training set filtering.

The experiments have been initially presented for Slovenian in our short paper in the 4REAL workshop (Repar et al. 2018). Here, we provide additional analysis and extend the experiments to the other two languages under consideration. The results are reported in Sect. 4.1 to 4.5.

In the 4REAL paper, precision was already relatively high (see for example line 8 in Table 3), which is why our additional experiments focused on improving recall. We implemented several additional approaches as reported in Sect. 4.6 to 4.8:

- Removing the Needleman–Wunsch Distance feature.
- Term length filtering.
- Adding new cognate-based features.

4.1 Giza++ terms only

We thought that one of the reasons for low results can be that not all EUROVOC terms actually appear in the Giza++ training data (i.e. DGT translation memory). The terms that do not appear in the Giza++ training data could have dictionary-based features similar to the generated negative examples, which could affect the precision of a classifier that was trained on those terms. We found that only 4,153 out of 7,083 Slovenian terms of the entire EUROVOC thesaurus do in fact appear in a DGT translation memory. Using only these terms in the classifier training set did provide modest improvements of precision, recall and F-score across all three languages. For details, see line 3 in Tables 3, 4 and 5.

4.2 Giza++ cleaning

The output of the Giza++ tool contained a lot of noise and we thought it could perhaps have a detrimental effect on the results. There is no mention of any sophisticated Giza++ dictionary cleaning in the original paper beyond removing all entries where probability is lower then 0.05 and entries where the source word is less than 4 characters and the target word more than 5 characters in length and vice versa (introduced to avoid stopword-content word pairs). For clean Giza++ dictionaries, we used the resources described in Aker et al. (2014), available via the META-SHARE repository[14] (Piperidis et al. 2014), specifically, the transliteration-based approach which yielded the best results according to the cited paper.

For Slovenian and Dutch, precision and F-score improved marginally at a cost of a lower recall, while for French, precision, recall and F-score all decreased. For details, see line 4 in Tables 3, 4 and 5.

4.3 Lemmatization

The original paper does not mention lemmatization which is why we assumed that all input data (Giza++ dictionaries, EUROVOC thesaurus) is not lemmatized. They

[14] http://metashare.tilde.com, last accessed: February 14, 2019.

state that to capture words with morphological differences, they don't perform direct string matching but utilize Levenshtein Distance and two words are considered equal if the Levenshtein Distance (Levenshtein 1966) is equal or higher than 0.95. This led us to believe that no lemmatization was used. Nevertheless, we thought lemmatizing the input data could potentially improve the results which is why we adapted the algorithm to perform lemmatization (using Lemmagen (Juršič et al. 2010)) of the Giza++ input data and the EUROVOC terms. We have also removed the Levenshtein distance string matching and replaced it with direct string matching (i.e. word A is equal to word B, if word A is exactly the same as B), which drastically improved the execution time of the software.

We considered lemmatization as a factor that could explain the difference in results obtained by us and Aker et al. (2013), but our experiments on lemmatized and unlemmatized clean Giza++ dictionaries show that lemmatization does not have a significant impact on the results. Compared to the configuration with unlemmatized clean Giza++ dictionaries, in the configuration with lemmatized Giza++ dictionaries precision was slightly lower (by 0.1%), recall was a bit higher (by around 4%) and F-score was lower by 0.2%. For details, see Table 3, line 4a. As lemmatization significantly slows down the experimentation, we tested the results first on Slovenian, where the influence of the lemmatization should be the largest as it is a morphologically-rich language. As lemmatization did not improve the results, we did not repeat the experiments for French and Dutch.

4.4 Changing the ratio of positive/negative examples in the training set

In the original paper, the training set is balanced (i.e. the ratio of positive vs. negative examples is 1) but the test set is not (the ratio is around 1:2000). Since our classifier had low precision and relatively high recall, we figured that an unbalanced training set with much more negative than positive examples could improve the former. To test this, we experimented with training the classifier on unbalanced train sets with different ratios between positive and negative examples. The general tendency we noticed during experimentation is that a very unbalanced train set (ratio of 1:200 between positive and negative examples[15]) greatly improves the precision of the classifier at a cost of somewhat lower recall, when compared to balanced train set or less unbalanced train set (e.g., ratio of 1:10 between positive and negative examples). For details, see line 5 in Tables 3, 4 and 5.

4.5 Training set filtering

The original paper mentions that their classifier initially achieved low precision on Lithuanian language training set, which they were able to improve by manually removing 467 positive term pairs that had the same characteristics as negative examples from the training set. No manual removal is mentioned for Slovenian, French and Dutch.

[15] 1:200 imbalance ratio was the largest imbalance we tried, since the testing results indicated that no further gains could be achieved by further increasing the imbalance.

We have performed an error analysis and found that many incorrectly classified term pairs are cases of partial translation where one unit in a multi-word term has a correct Giza++ dictionary translation in the corresponding term in the other language. Some EN-SL examples can be seen in Table 6, and similar errors were observed for for the other two language pairs.

Based on this problem of partial translations, leading to false positive examples, we focused on the features that would eliminate these partial translations from the training set. After a systematic experimentation, we noticed that we can drastically improve precision if we only keep positive term pairs with the following feature values in the training set:

- isfirstwordTranslated = True.
- islasttwordTranslated = True.
- percentageOfCoverage > 0.66.
- isfirstwordTranslated-reversed = True.
- islasttwordTranslated-reversed = True.
- percentageOfCoverage-reversed > 0.66.

Using this approach, we managed to greatly increase precision at a cost of significant drop in recall values for all three languages. For details see line 6 (*Training set filtering 1*) in Tables 3, 4 and 5. When combining this approach with an unbalanced dataset described in the previous section, we managed to improve precision even further, but again at a cost of lower recall. For details, see lines 7 and 8 (*Training set filtering 2 and 3*) in Tables 3, 4 and 5.

4.6 Cognate feature analysis and removing the Needleman–Wunsch Distance feature

We performed an analysis of the results on the English–Slovenian language pair achieved with the best configuration for precision (line 8—*Training set filtering 3* in Table 3) in our experiments (Repar et al. 2018) and discovered that cognate term pairs were not being considered by the classifier. In a way, this was expected since in the previous step we have filtered the training set based on mostly dictionary-based features.

When analyzing the performance of the cognate-based features, we found that four (Longest Common Subsequence Ratio (LCSSR) Longest Common Substring Ratio (LCSTR), Dice Similarity (Dice), Normalized Levenshtein Distance (nLD)) out of five perform as expected with cognate term pairs having high values, but Needleman-Wunsch Distance (NWD) did not. As already mentioned in the beginning, the formula provided by the authors for computing NWD feature possibly contained an error, therefore we opted for the implementation as mentioned in Nassirudin and Purwarianti (2015). Table 7 shows the behaviour of the five cognate-based features. When we are dealing with actual cognates, all five features have high values, but when the two terms in questions are not cognates, only NWD stays high.

Table 6 Examples of negative term pairs misclassified as positive

EN	SL	Giza++
Agrarian reform	Kmetijski odpadki	Agrarian, kmetijske, 0.29737
Brussels region	Območje proste trgovine	Region, območje, 0.0970153
Energy transport	Nacionalni prevoz	Transport, prevoz, 0.442456
Fishery product	Tekstilni izdelek	Product, izdelek, 0.306948

Column 1 contains the English term, column 2 contains the Slovenian term and column 3 contains the Giza++ dictionary entry (from the non-clean version, see Sect. 4.2) responsible for positive dictionary-based features

Table 7 Cognate-based features values (showing issues with NWD)

EN	SL	LCSSR	LCSTR	Dice	nLD	NWD
hospitalisation	hospitalizacija	0.73	0.60	0.60	0.73	0.6
monopsony	monopson	0.89	0.89	0.94	0.89	1.00
fish	predstavniška demokracija	0.12	0.12	0.20	0.12	0.75
Yemen	osna obremenitev	0.25	0.25	0.38	0.25	0.80

The first two term pairs are actual cognates with all five cognate-based features having high values. The last two pairs are not cognates and show the issues with the Needleman-Wunsch Distance (NWD), which is the only measure that keeps a high value. Note that due to character mapping rules (see Section 3.2.1.), the word "predstavniška" was transformed into "predstavnishka"

For this reason, we ran our experiments without the NWD feature, but the results did not improve since the SVM classifier is known to be capable of handling noisy features.

4.7 Term length filtering

Based on error analysis, one of the major issues confusing the classifier were training examples with differing word lengths. E.g., the source term in the example would have one word, but the target term would have two. An analysis of the terms in Eurovoc for the three language pairs in question showed that 26% of the EN-SL term pairs, 34% of the EN-FR term pairs and 48% of the EN-NL term pairs have different word lengths of the source and target terms (the reason for the high ratio in EN-NL is the use of compounds in Dutch). This turned out to be one of the characteristics leading to low classification performance: for Slovenian with the replicated configuration (line 2 in Table 3) the classifier returned a total of 14,721 positively classified examples. 14,193 out of these were false positives—incorrectly aligned term pairs. A further 13376 out of these had different lengths of the source and target terms. A visual inspection of feature values indicated that there is often no clear difference between positive and negative term pairs (see Table 8).

Since this was an issue, we experimented with additional term length filtering. We took the positively classified examples from the *training set filtering 1* approach as described in Sect. 4.5 (see line 6 in the tables) and added an additional filter: if the two terms do not have the same number of words, we change the prediction from positive to negative. Using this additional filter, we achieved good precision for Slovenian (81%), and respectable for French (68%) and Dutch (76%). On the other hand, recall values were badly affected, since one third of positive term pairs in the constructed test set are terms of different word length (meaning that highest possible theoretical recall with this approach is 66%). Recall was again best for Slovenian with a value close to 50% and a bit worse for French and Dutch with a value at around 40% and 37% respectively. Consequently, F-scores were the highest for Slovenian and lower for Dutch and French. For details, see line 9 in Tables 3, 4 and 5.

From the original paper it is clear, that authors were aware of the possible complexity of terms of unequal length, as they consider terms of different lengths in the test set construction. So, we exclude the possibility that authors did not have such examples in the test set.

4.8 Cognate-based feature approach

The analysis showed that all *Training set filtering* approaches tend to overestimate the importance of Giza++ features and underestimate cognate-based features. This results in a low recall for correct cognate term pairs, which are rarely classified as positive, if their Giza++ based feature values do not show similarity with Giza++ based feature values for non-cognate correct term pairs. For example, Giza++ dictionary does not contain a Slovenian translation *pacifizem* for the English term *pacifism*, which means that the values of features *isFirstWordTranslated*, *isLastWordTranslated*, *isFirstWordTranslated-reversed* and *isLastWordTranslated-reversed* are False and the values for features *percentageOfCoverage* and *percentageOfCoverage-reversed* are zero, therefore the classifier would have a strong inclination to classify this correct term pair as incorrect, even though cognate based feature values clearly indicate that these two terms are cognates.

In order to improve the detection of cognate terms, we first propose two new cognate based features:

- isFirstWordCognate: a binary feature which returns True if the longest common consecutive string (LCST) of the first words in the source and target terms divided by the length of the longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters. For example, the value of the feature for the English-Slovenian term pair *Klaipeda county - Klaipedsko okrožje* would be True because the LCST for the first words in both terms is *Klaiped*, which has a length of 7. The length of the longest of the two first words in the terms (*Klaipedsko*) is 10 and 7 divided by 10 is 0.7, which is equal to the threshold value.
- isLastWordCognate: a binary feature which returns True if the longest common consecutive string (LCST) of the last words in the source and target terms

Table 8 A comparison of dictionary feature values

Source term	raw material	provision	additional resources	provision
Target term	surovine	računovodska rezervacija	surovine	urbanistični predpisi
Correctly aligned	True	True	False	False
isFirstWordTranslated	1	0	0	0
isLastWordTranslated	1	1	1	1
pctOfTransWords	0.5	1	0.5	1
pctOfNotTransWords	0.5	0	0.5	0
longestTransUnitInPct	0.5	1	0.5	1
longestNotTransUnitInPct	0.5	0	0.5	0
isFirstWordTranslated_R	0	0	0	0
isLastWordTranslated_R	1	1	1	1
pctOfTransWords_R	1	0.5	1	0.5
pctOfNotTransWords_R	0	0.5	0	0.5
longestTransUnitInPct_R	1	0.5	1	0.5
longestNotTransUnitInPct_R	0	0.5	0	0.5

The first two term pairs are correctly alligned term pairs from the training set (line 2 in Table 3), the second two are not correctly alligned term pairs. We can observe that the dictionary feature values are very similar—compare for example *raw material/surovine* and *additional resources/surovine*

divided by the length of longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters. For example, the value of the feature for the English-Slovenian term pair *Latin America - Latinska Amerika* would be True because the LCST for the last words in both terms is *Ameri*, which has a length of 5. The length of the longest of the two last words in the terms is 7 and 7 divided by 5 is 0.714, which is greater than the threshold value.

As having the same number of words in the source and target term could play a role in classification, we also add three new features responsible for encoding term length information:

- sourceTargetLengthMatch: a binary feature that returns True if the number of words in source and target terms match.
- sourceTermLength: returns the number of words in the source term.
- targetTermLength: returns the number of words in the target term.

Analysis of the filtered training set showed that it contained a small number of positive cognate based term pair examples, therefore the first step was to include more of them into the dataset. We build three separate datasets, each of them filtered according to the following feature values:

- isFirstWordCognate = True and isLastWordCognate = True.
- isFirstWordTranslated = True and isLastWordCognate = True.
- isFirstWordCognate = True and isLastWordTranslated = True.

The terms from these three datasets are added to the original filtered train set (we make sure that each positive term pair is represented in the new dataset only once by removing all the duplicates). The new dataset contains two distinct groups of terms, one with favorable Giza++ based features (and unfavorable cognate based features) and one with favorable cognate based features (and in some cases unfavorable Giza++ based features). Since this new dataset structure represents a classic "exclusive or" (XOR) problem which a linear classifier is unable to solve, we also replace the linear kernel of our SVM classifier with the Gaussian one.

Using this approach, precision was close to 90% (Slovenian, French) or just over 90% (Dutch), recall was just over 50% for Slovenian, around 52% for Dutch and close to 40% for French. For details, see line 10 in Tables 3, 4 and 5.

4.9 Best results

Overall, the setting with the best precision is Train set filtering 3. Compared to the replicated approach (line 2 in Tables 3, 4 and 5), it has an unbalanced dataset of 1:200 (see Section 4.4) and employs the term filtering strategy described in Sect. 4.5. However, for a small gain in recall at the price of a slight decrease in precision, a good alternative is the Cognates approach (line 10 in Tables 3, 4 and 5), which is

based on the Train set filtering 3 approach and additionally includes the cognate detection strategies described in Sect. 4.8.

5 Manual evaluation

The first part of this section contains the manual evaluation replicated from Aker et al. (2013), already reported in Repar et al. (2018), while the second part is novel and contains an evaluation using a new dataset and has a specific focus on cognate term pairs.

5.1 Replicating the manual evaluation experiments from the original paper

Similar to the original paper, we also performed manual evaluation. We selected a random subset of term pairs classified as positive by the classifier (using the *Training set filtering 3* configuration (line 8 in Table 3) that yielded the best precision). While the authors of the original approach extract monolingual terms using the term extraction and tagging tool TWSC (Pinnis et al. 2012), we use a workflow for monolingual term extraction by Pollak et al. (2012). Both use a similar approach - terms are first extracted using morphosyntactic patterns and then filtered using statistical measures: TWSC uses pointwise mutual information and TF*IDF, while Pollak et al. (2012) is based on an approach by Vintar (2010) and compares the relative frequencies of words composing a term in the domain-specific (i.e. the one we are extracting terminology from) corpus and a general language corpus.

In contrast to the original paper where they extracted terms from domain-specific Wikipedia articles (for the English-German language pair), we are using two translation memories—one containing finance-related content, the other containing IT content. Another difference is that extraction in the original paper was done on comparable corpora, but we extracted terms from parallel corpora - which is why we expected our results to be better. Each source term is paired with each target term (just as in the original paper - if both term lists contained 100 terms, we would have 10,000 term pairs) and extract the features for each term pair. The term pairs were then presented to the classifier that labeled them as correct or incorrect term translations. Afterwards, we took a random subset of 200 term pairs classified as correct and showed them to an experienced translator[16] fluent in both languages who evaluated them according to the criteria set out in the original paper:

- **1—Equivalence**: The terms are exact translations/transliterations of each other (e.g., *type—tip*).
- **2—Inclusion**: Not an exact translation/transliteration, but an exact translation/transliteration of one term is entirely contained within the term in the other language (e.g., *end date—datum*).
- **3—Overlap**: Not category 1 or 2, but the terms share at least one translated/transliterated word (e.g., *user id—uporabniško ime*).

[16] The original paper used two annotators, hence two lines for each domain in Table 4.

– **4—Unrelated**: No word in either term is a translation/transliteration of a word in the other (e.g., *level—uporabnik*[17]).

The results of the manual evaluation can be found in Table 9. Manual evaluation showed that 72% of positive term pairs in the Finance domain, and 79% of positive term pairs in the IT domain were correctly classified by the classifier. The differences between the *Finance* and *IT* datasets can be partially explained by the *Finance* dataset containing more MWE terms than the *IT* dataset (84 vs. 51 for SL and 78 vs. 49 for EN). On the one hand, this means that the chances of aligning a single word term in one language with a multi word term in another language is greater, hence the greater number of partial translations in *Finance* (category 2 - Inclusion), while on the other, single word terms means less characters for the algorithm to work with, hence the greater number of outright mistakes in *IT* (category 4 - Unrelated). Compared to the original paper, we believe these results are comparable when taking into account the different monolingual extraction procedures , the different language pairs and the human factor related to different annotators.

5.2 Evaluation on a Karst terminology gold-standard

As mentioned in Sect. 4, the best configuration in terms of precision used in Repar et al. (2018) (line 8 in Tables 3, 4 and 5) overestimates dictionary-based and underestimates cognate-based features. To alleviate this, we added additional features and filtering strategies to our approach to try to improve cognate term pair alignment (see lines 9 and 10 in the results tables). However, evaluating its performance on EUROVOC is difficult as many terms have favorable dictionary-based features due to the fact that both the Giza++ dictionary and EUROVOC are made from the same content (i.e. EU documentation). For the evaluation in this section, we therefore selected a domain, with a content type which is unlikely to be found in DGT (Steinberger et al. 2013), i.e. karstology, which is the science in the field of geomorphology, specializing in the study of karst formations.

To evaluate our bilingual term alignment approach, we used a gold standard of EN-SL aligned karst terminology,[18] which was manually created by the authors of the karstology corpus (Vintar and Grčić-Simeunović 2016). The gold standard consists of 52 English-Slovenian term pairs. For the evaluation experiment, we aligned all Slovenian term with all English terms, resulting in a dataset of 52 positive examples and 2652 negative examples. With the best configuration for precision (line 8 in Table 3), selected also as the best configuration in Repar et al. (2018), precision was 100%, but recall was only 40.4%. Many term pairs containing cognates such as “eogenetic cave—eogenetska jama”, “epigenic aquifer—epigeni vodonosnik” or “karst polje—kraško polje”, were not aligned. With the final cognate approach (line 10 in Table 3), we managed to retain 100% precision and raise the recall to 50% by finding 7 additional cognate term pairs (*aggressive*

[17] “uporabnik” means “user”.

[18] http://source.ijs.si/mmartinc/4real2018/tree/master/datasets/karst_corpus.

water—agresivna voda, *eogenetic cave—eogenetska jama*, *precipitation—precipitacija*, *ponor cave—ponorna jama*, *epigenic aquifier—epigeni vodonosnik*, *karst polje—kraško polje*, *linear stream cave—linearna epifreatična jama*). However one half of correct term pairs remain undiscovered. We believe this is due to 1) domain-specific words which are not cognates and are missing from the Giza++ dictionary (e.g., *porous aquifer—medzrnski vodonosnik* and *denuded cave—brezstropa jama*), and 2) valid cognate words which do not meet the threshold described in Sect. 4.8 (e.g. *oxidization—oksidacija*, *percolation—perkolacija* and *liquefication—likvifakcija*).[19]

6 Replicability and reproducibility of our own terminology alignment results

As mentioned before, availability of the source code can drastically improve the reproducibility of experiments, since very detailed descriptions of procedures used in the experiments are beyond the scope of most papers because of length limitations and negative effects on the readability of the paper. Since we wanted to ensure the full reproducibility of our approach, we decided to publish the source code for all the conducted experiments and results that are published in the paper. As we were aware that just the presence of source code itself does not guarantee complete reproducibility, we decided that the published code should comply to the following three criteria:

- Instructions on how to use the code should be as unambiguous, simple and clear as possible.
- Code should be bug free and running it according to the instructions should yield the exact same results as published in the paper.
- Running the code should require as little time and technical skills as possible.

In order to validate that the published code complies to these criteria, we asked three students[20] to try to reproduce the results published in the paper (Repar et al. 2018) and after that answer the following questions related to the proposed criteria:

- Did you manage to reproduce the results?
- If not, what do you think was the main problem?
- If yes, how much time did you need for replicating the experiment?
- Were the instructions clear?
- Did you run into any specific problems during any part of the replicability attempt? If yes, please describe it.
- Do you have any suggestions on how to further improve the reproducibility of the results?

[19] It might also make sense to include morphological information as a feature of the machine learning algorithm, since all these word have endings typical of cognates in their respective languages.

[20] 2 Master students (one in Economy and one in Computer science) and 1 first year PhD student in ICT.

Table 9 Manual evaluation results

Domain	1	2	3	4
Reported in Aker et al. (2013)				
IT, Ann. 1	0.81	0.06	0.06	0.07
IT, Ann. 2	0.83	0.07	0.07	0.03
Auto, Ann. 1	0.66	0.12	0.16	0.06
Auto, Ann. 2	0.60	0.15	0.16	0.09
Replication				
Finance	0.72	0.09	0.12	0.07
IT	0.79	0.01	0.09	0.12

Ann. stands for "Annotator" since the original paper uses two annotators

We also imposed a time limit of 8 hours (one working day) for the entire replicability attempt. If that limit was reached, the replicability attempt would count as unsuccessful.

The feedback we got was interesting and made us reconsider the initial source code criteria. Two out of three students managed to reproduce all the published results in less than an hour without any major problems. They did however point out some mistakes and ambiguities in the instructions on how to run the code. These were mostly connected with the programming environment used by the students, one of them using PyPI Python package manager for acquiring dependencies while the other one used the Conda environment, for which the usage instructions were not published.

The third student managed to reproduce the results in about two hours and reported some major problems with dependencies installation. He was the only person trying to reproduce the experiments in the Windows environment while the other two students used a Linux operating system, and he reported problems with the Python implementation of the Lemmagen lemmatizer (Juršič et al. 2010), which he was unable to install properly on the Windows platform. He managed to overcome the problem by manually removing the dependency from the code, by which he limited the flexibility of the published source code (he could only use it for the classification on the pre-generated train and test sets) but did not make the reproduction impossible.

While he was successful at reproducing the results for eight out of nine experiments published in the paper, he also reported a slight deviation (by less than 0.05 percentage point) from the reported recall and F-score in one of the experiments. Although we are not sure what is the exact reason for this deviation, we suspect it could be connected to the difference in operating systems.

These experiments show that programming environment and the choice of the operating system can have an unexpected negative impact on the reproducibility. While attaching code usage instructions for every possible programming environment and operating system is practically impossible, we do believe that the results of this experiment show that a published source code should comply to one additional criteria:

– Instructions should clearly specify on which operating system and in which programming environment the reported results were produced.

We have updated the usage instructions for our source code to comply with these criteria.

7 Reusability of our code in the ClowdFlows online platform

Because we want to make sure that our terminology alignment system is also available to a wider audience of users with lower level of technical skills (e.g., translators or linguists) and because we want to encourage a very simple reusability of our system, we have implemented the system into a cloud-based visual programming platform ClowdFlows (Kranjc et al. 2012). The ClowdFlows platform employs a visual programming paradigm in order to simplify the representation of complex data mining procedures into visual arrangements of their building blocks. Its graphical user interface is designed to enable the users to connect processing components (i.e. widgets) into executable pipelines (i.e. workflows) on a design canvas by a drag and drop technique, reducing the complexity of composition and execution of these workflows. The platform also enables online sharing of the composed workflows.

We took pretrained models of our terminology alignment system for English-Slovenian, English-French and English-Dutch alignment and packed them in a widget *Terminology alignment*, so it can be used out-of-the-box. The widget takes two columns of the Pandas dataframe (McKinney 2011) containing the source and target terms as inputs and returns a dataframe containing aligned term pairs. The user needs to define the names of the columns in the dataframe containing source and target language termlists, and the language of alignment as parameters. The user can also switch between configurations *Training set filtering 3* with the best precision and *Cognates approach* with the on average best F-score for all three languages while still having good precision by either enabling or disabling the *Maximize recall* widget parameter. Such an end to end system for bilingual terminology alignment in ClowdFlows is implemented at: http://clowdflows.org/workflow/13789/.[21] Another widget called *Terminology alignment evaluation* is used for determining the performance of the system (if we have a gold standard available), taking as input the dataframe produced by the *Terminology alignment* widget and a dataframe containing true alignments, and outputting the performance score in terms of precision, recall and F-score.

Workflow in Fig. 1 (available at http://clowdflows.org/workflow/13753/) is a ClowdFlows implementation for terminology alignment and evaluation. The source and target terminologies are both loaded from a CSV file with the help of the *Load Corpus From CSV* widget and fed as input to the *Terminology alignment* widget,

[21] Note that the execution time of term alignment increases rapidly with the increase in number of terms, e.g., alignment of hundred terms takes around five minutes, while it takes about one hour for alignment of thousand terms.

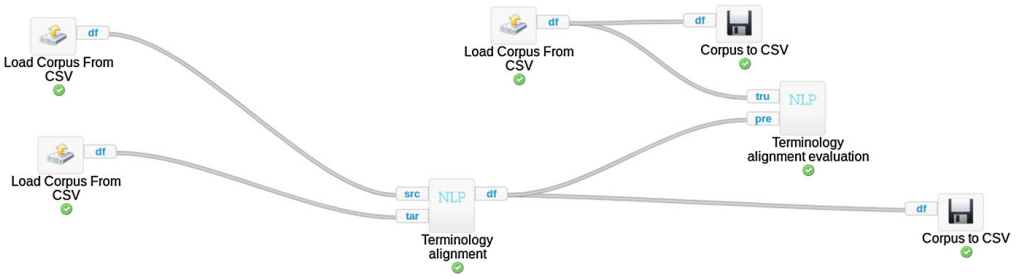

Fig. 1 ClowdFlows implementation of the system for terminology alignment and evaluation available at http://clowdflows.org/workflow/13753/

which returns a dataframe with alignments. These are written to a CSV file with the *Corpus to CSV* widget and also fed to the *Terminology alignment evaluation* widget together with the dataframe containing true alignments (which was also loaded from a CSV file with the *Load Corpus From CSV* widget) in order to estimate the performance of the system. In addition, term alignment widget can also be incorporated into a bilingual terminology extraction workflow (Pollak et al. 2012). The workflow with the newly added term alignment widget, is available at http://clowdflows.org/workflow/13723/), where a user can now input text from a specific domain in Slovenian and English and get aligned terminology as output.

8 Conclusions and future work

Based on our research and attempts at replicating a bilingual terminology alignment paper reproducing its results, we propose a set of best practices any bilingual terminology extraction paper (and more generally every NLP paper) should fulfill to facilitate reproducibility and replicability of the experiments:

- *Dataset availability.* Availability of datasets (i.e. gold standard term lists, corpora) is an essential prerequisite for successful replication.
- *Experiment code availability.* The main task of reproducibility and replicability experiments is often to reconstruct the experiments in computer code. It is a cumbersome process which inevitably requires that the reproducer/replicator makes educated guesses at some point since a detailed description of the code is beyond the scope of most papers. Having the original code available greatly increases the ease of reproducibility and replicability experiments.
- *Tool availability.* Availability of a tool or application (online or offline) where experiments can be conducted eases reproducibility and replicability, but also enables the reusability of results by a larger community.
- Finally, releasing intermediate results, configuration settings and the actual outcomes of individual experiments, while not essential, would provide future researchers with an even greater possibility of successful reproduction of the paper's results.

A prerequisite for successful reproduction and replication is a clearly written research paper. However as is evident from our example, it is often difficult to include all necessary implementation notes given the length restrictions of the paper. For this reason, another best practice would be to provide relevant implementation examples alongside the code (which is what we did for feature construction.[22]) Finally, as the experiment in Sect. 6 showed, even code itself is sometimes not enough without additional implementation notes and information on the operating systems and software used. In addition, testing the code by non-authors is strongly recommended.

Our attempts focused on the approach to bilingual term alignment using machine learning by Aker et al. (2013). They approach term alignment as a bilingual classification task—for each term pair, they create various features based on word dictionaries (i.e. created with Giza++ from the DGT translation memory) and word similarities across languages. They evaluated their classifier on a held-out set of term pairs and additionally by manual evaluation. Their results on the held-out set were excellent, with 100% precision and 66% recall for the English-Slovenian and English-French language pair and 98% precision and 82% recall for English-Dutch.

Our reproduction attempt focused on three language pairs: English-Slovenian, English-Dutch and English-French (in contrast with the original article where they had altogether 20 language pairs) and we were unable to reproduce the results following the procedures described in the paper. In fact, our results have been dramatically different from the original paper with precision being less than 4% and recall close to 90% for all three language pairs under consideration. We then tested several different strategies for improving the results ranging from Giza++ dictionary cleaning, lemmatization, different ratios of positive and negative examples in the training and test sets, training set filtering based on feature values and term length, and adding new cognate-based features. The most effective strategies employed unbalanced training set and training set filtering based on certain feature values which resulted in precision exceeding 90% for all three language combinations (*Training set filtering 3* configuration, line 8 in Tables 3, 4 and 5). It is possible that in the original experiments authors performed a similar training set filtering strategy, because the original paper mentions that their classifier initially achieved low precision on Lithuanian language training set, which they were able to improve by manually removing positive term pairs that had the same characteristics as negative examples from the training set. However, no manual removal is mentioned for Slovenian, Dutch or French. Further attempts were directed at boosting recall and the performance of cognate-based features. By adding additional cognate-based features, we were able to improve recall by around 16% for Dutch, 8% for French and by around 2% for Slovenian (over the *Training set filtering 3* configuration) at a cost of a moderate drop in precision.

For evaluation we focused only on Slovenian, which is our native language and of primarily interest for our applied tasks. We performed manual evaluation similar to the original paper and reached roughly the same results with our adapted approach. In addition, because we discovered that Eurovoc data is of limited use for

[22] http://source.ijs.si/mmartinc/4real2018/blob/master/feature_examples.docx.

evaluating the performance of cognate-based features, we ran experiments on an English-Slovenian karstology gold standard term list. With the *Cognates approach* configuration (line 10 in Tables 3, 4 and 5), we improved recall by 11% (compared to the *Training set filtering 3* configuration) and a qualititive analysis of the results showed that the new strategies for boosting the performance of cognate-based features do indeed result in more cognate term pairs being properly aligned.

This paper demonstrates some of the obstacles for research reproducibility and replicability, with the prime one being code unavailability. Had we had access to the code of the original experiments, it is highly likely that replicating the original paper would be a trivial matter. Also in this particular case, the discrepancy in the results could be attributed to the scope of the original paper - with more than 20 languages—which is also a demonstration of very impressive approach—it would be impossible to describe procedures for all of them. We weren't able to reproduce the results of the original paper, but after developing the optimization approaches described above over the course of several months, we were able to reach a useful outcome at the end. We believe that providing supplementary material online, i.e. the code and datasets, is the only way of assuring complete reproducibility of results. For this reason, in order to help with any future reproducibility/replicability attempts of our paper, we are publishing the code at: http://source.ijs.si/mmartinc/4real2018.

In terms of future work, we plan to expand the feature set by introducing the features derived from the distributions in parallel corpora (e.g. co-frequency, logDice and other measures, see Baisa et al. (2015)), as well as investigate novel methods using cross-lingual embeddings. In terms of reproducibility, we plan to extend the study to a systematic comparison of different term alignment and term extraction methods.

Acknowledgements This paper is supported by European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The authors acknowledge also the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103). The authors also acknowledge the project TermFrame—Terminology and Knowledge Frames across Languages (No. J6-9372), which was financially supported by the Slovenian Research Agency. We would also like to thank the company Iolar, for allowing us to use the data from the translation memories in one of the experiments.

References

Aker, A., Paramita, M., & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Vol. 1*. Long Papers (pp 402–411).

Aker, A., Paramita, M. L., Pinnis, M., & Gaizauskas, R. (2014). Bilingual dictionaries for all EU languages. In *Proceedings of 9th International Conference on Language Resources and Evaluation*. (pp 2839–2845).

Arčan, M., Turchi, M., Tonelli, S., & Buitelaar, P. (2014). Enhancing statistical machine translation with bilingual terminology in a CAT environment. https://doi.org/10.13140/2.1.1019.8404.

Bader, B. W., & Chew. P. A. (2008). Enhancing multilingual latent semantic analysis with term alignment information. In *Proceedings of the 22nd International Conference on Computational Linguistics: Vol. 1. Association for Computational Linguistics* (pp 49–56).

Baisa, V., Ulipová, B., & Cukr, M. (2015). Bilingual terminology extraction in Sketch Engine. In *9th Workshop on Recent Advances in Slavonic Natural Language Processing*. (pp 61–67).

Baldwin, T., & Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. (pp 24–31).

Bouamor, D., Semmar, N., Zweigenbaum, P. (2013). Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Vol. 2: Short Papers*. (pp 759–764).

Branco, A., Calzolari, N., & Choukri, K. (eds) (2018). 4REAL 2018—Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language, ELRA.

Branco, A., Cohen, K. B., Vossen, P., Ide, N., & Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: introducing an LRE special section.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Cao, Y,, & Li, H. (2002). Base noun phrase translation using web data and the EM algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics: Vol. 1*. (pp 1–7).

Chiao, Y. C., & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics: Vol. 2*. (pp 1–5).

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T., Goss, F., Ide, N., Névéol, A., Grouin, C., & Hunter, L. E. (2018). Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. (pp 156–165).

Daille, B., Gaussier, E., & Langé, J. M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics: Vol. 1*. (pp 515–521).

Daille, B., & Morin, E. (2005). French-English terminology extraction from comparable corpora. *Natural Language Processing - IJCNLP*, *2005*, 707–718.

Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P., & Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Vol. 1: Long Papers*. (pp 1691–1701).

Foo, J. (2012). Computational terminology: Exploring bilingual and monolingual term extraction. PhD thesis, Linköping University Electronic Press.

Fung, P., & Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics: Vol. 1*. (pp 414–420).

Gaizauskas, R., Aker, A., & Yang Feng, R. (2012). Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics*. (pp 23–32).

Gale, W., & Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, *19*(1), 75–102.

Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th International Conference on Computational Linguistics: Vol. 1*. (pp 444–450).

Ha, L. A., Fern, G., Mitkov, R., Corpas, G. (2008). Mutual bilingual terminology extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. (pp 1818–1824).

Haque, R., Penkale, S., & Way, A. (2014). Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. (pp 42–51).

Hazem, A., & Morin, E. (2016). Efficient data selection for bilingual terminology extraction from comparable corpora. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. (pp 3401–3411).

Hazem, A., & Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the 8th International Joint Conference on Natural Language Processing: Vol. 1: Long Papers*. (pp 685–693).

Ideue, M., Yamamoto, K., Utiyama, M., & Sumita., E. (2011). A comparison of unsupervised bilingual term extraction methods using phrase tables. In *Proceedings of the 13th Machine Translation Summit*. (pp 346–351).

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Alphen aan den Rijn: Kluwer Academic Publishers.

Johnson, I., & Macphail, A. (2000). IATE–Inter-Agency Terminology Exchange: Development of a single central terminology database for the institutions and agencies of the European Union. In *Proceedings of the Workshop on Terminology resources and computation, LREC 2000 Conference*.

Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, *16*(9), 1190–1214.

Kano, Y., Baumgartner, W. A, Jr., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L., et al. (2009). U-compare: Share and compare text mining tools with uima. *Bioinformatics*, *25*(15), 1997–1998.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit: Vol. 5*. (pp 79–86).

Kontonatsios, G., Korkontzelos, I., Tsujii, J., & Ananiadou, S. (2014). Combining string and context similarity for bilingual term alignment from comparable corpora. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (pp 1701–1712).

Kranjc, J., Podpečan, V., & Lavrač, N. (2012). Clowdflows: A cloud based scientific workflow platform. In *Proceedings of Machine Learning and Knowledge Discovery in Databases, ECML/PKDD (2)*. Springer. (pp 816–819).

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*. (pp 17–22).

Lee, L., Aw, A., Zhang, M., & Li, H. (2010). Em-based hybrid model for bilingual terminology extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics*. (pp 639–646).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707.

Macken, L., Lefever, E., & Hoste, V. (2013). Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology International Journal of Theoretical and Applied Issues in Specialized Communication*, *19*(1), 1–30.

McKinney, W. (2011). Pandas: A foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing. (pp 1–9).

Morin, E., Daille, B., Takeuchi, K., Kageura, K. (2007). Bilingual terminology mining-using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. (pp 664–671).

Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2008). Brains, not brawn: The use of smart comparable corpora in bilingual terminology mining. *ACM Transactions on Speech and Language Processing*, *7*(1), 1.

Nassirudin, M., & Purwarianti, A. (2015). Indonesian-Japanese term extraction from bilingual corpora using machine learning. In *International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2015*. (pp 111–116).

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Pedersen, T. (2008). Empiricism is not a matter of faith. *Computational Linguistics*, *34*(3), 465–470.

Pinnis, M., Ljubešić, N., Stefanescu, D., Skadina, I., Tadić, M., & Gornostaya, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*. (pp 20–21).

Piperidis, S., Papageorgiou, H., Spurk, C., Rehm, G., Choukri, K., Hamon, O., Calzolari, N., Del Gratta, R., Magnini, B., & Girardi, C. (2014). Meta-share: One year after. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. (pp 1532–1538).

Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N., & Vintar V,. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. *11th Conference on Natural Language Processing, KONVENS 2012 - Empirical Methods in Natural Language Processing* (pp. 53–60). Vienna: Austria.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712.

Rapp, R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. (pp 519–526).

Repar, A., Martinc, M., & Pollak, S. (2018). Machine learning approach to bilingual terminology alignment: Reimplementation and adaptation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P. (2013). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.

Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. Computational Linguistics and Intelligent Text Processing. (pp 101–121).

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In: Chair) NCC, Choukri K, Declerck T, Dogan MU, Maegaard B, Mariani J, Odijk J, Piperidis S (eds) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey.

Vintar, Š., & Grčić-Simeunović (2016). Definition frames as language-dependent models of knowledge transfer. Fachsprache : internationale Zeitschrift für Fachsprachenforschung, - didaktik und Terminologie. (pp 43–58).

Vintar, Š. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology International Journal of Theoretical and Applied Issues in Specialized Communication*, *16*(2), 141–158.

Wieling, M., Rawee, J., & van Noord, G. (2018). Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, *44*(4), 641–649.

Yentis, S., Campbell, F., & Lerman, J. (1993). Publication of abstracts presented at anaesthesia meetings. *Canadian Journal of Anaesthesia*, *40*(7), 632–634.

3.3 Classification with Word Embeddings, Dictionary and Cognate-based features

3.3.1 Description of the approach

Next, we focused on implementing cross-lingual word-embeddings into our approach. In Repar et al., 2021, we used VecMap by Artetxe et al. (2018) to map the Fasttext (Bojanowski et al., 2016) embeddings between Slovenian and English and then for every word calculate the top 3 closest word embeddings in the opposite language based on cosine distance. These alignments were then used as features for the machine learning algorithm in addition to the cognate and dictionary-based features. Using embedding-based features as a replacement for dictionary-based features provides a significant benefit: while a large bilingual parallel corpus is required to generate the Giza++ word alignment lists, no such data is required for embedding-based features where the only required inputs are two unrelated monolingual corpora and a small bilingual dictionary from which the embedding alignments are calculated.

A similar approach was used in Ulčar et al. (2021) in the section on terminology alignment, but instead of Fasttext, ELMo embeddings (Peters et al., 2018) were used with different languages and different mapping methods (Vecmap (Artetxe et al., 2018) and MUSE (Conneau et al., 2018), and ELMoGAN-O and ELMoGAN-10k (Ulčar et al., 2021)).

3.3.2 Results

The results were a mixed bag when using all available features. Without any training set filtering, the new features improve precision at the expense of recall, but are less effective when filtering is applied. Nevertheless, when we use additional trainset filters for the cognates approach, we observed a slight increase in both precision and recall resulting in the overall highest F_1 score. When we used only embedding-based and cognate-based features, which would be beneficial for language pairs without access to large parallel corpora needed to create Giza++ word alignments, there was a significant drop in recall in all cases, but precision actually increased when trainset filtering is applied and the cognates approach achieved the overall best precision.

Word-embedding based bilingual terminology alignment

Andraž Repar[1], Matej Martinc[2], Matej Ulčar[3], Senja Pollak[4]

[1]International Postgraduate School, Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia
[2] Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia
[3] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, Ljubljana, Slovenia
[4] Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia
E-mail: andraz.repar@ijs.si, matej.martinc@ijs.si, matej.ulcar@fri.uni-lj.si, senja.pollak@ijs.si

Abstract

The ability to accurately align concepts between languages can provide significant benefits in many practical applications. In this paper, we extend a machine learning approach using dictionary and cognate-based features with novel cross-lingual embedding features using pretrained fastText embeddings. We use the tool VecMap to align the embeddings between Slovenian and English and then for every word calculate the top 3 closest word embeddings in the opposite language based on cosine distance. These alignments are then used as features for the machine learning algorithm. With one configuration of the input parameters, we managed to improve the overall F-score compared to previous work, while another configuration yielded improved precision (96%) at a cost of lower recall. Using embedding-based features as a replacement for dictionary-based features provides a significant benefit: while a large bilingual parallel corpus is required to generate the Giza++ word alignment lists, no such data is required for embedding-based features where the only required inputs are two unrelated monolingual corpora and a small bilingual dictionary from which the embedding alignments are calculated.

Keywords: terminology alignment; word embeddings; embeddings alignment; machine learning

1. Introduction

The ability to accurately align concepts between languages can provide significant benefits in many practical applications. For example, in terminology, terms can be aligned between languages to provide bilingual terminological resources, while in the news industry, keywords can be aligned to provide better news clustering or search in another language. Accurate bilingual resources can also serve as seed data for various other NLP tasks, such as multilingual vector space alignment.

Bilingual terminology alignment[1] is the process of aligning terms between two candidate term lists in two languages. The primary purpose of bilingual terminology extraction is to build a term bank - i.e. a list of terms in one language along with their equivalents in the other language. With regard to the input text, we can distinguish between alignment on the basis of a parallel corpus and alignment on the basis of a comparable corpus. For the translation industry, bilingual terminology extraction from parallel corpora is extremely relevant due to the large amounts of sentence-aligned parallel corpora available in the form of translation memories. Consequently, initial attempts at bilingual terminology extraction involved parallel input data (Kupiec, 1993; Daille et al., 1994; Gaussier, 1998), and the interest of the community continued until today. However, most parallel corpora are owned by private companies[2], such as language service providers, who consider them to be their intellectual property and are reluctant to share them publicly. For this reason (and in particular for language pairs not involving English) considerable efforts have also been invested into researching bilingual terminology extraction from comparable corpora (Fung & Yee, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002; Cao & Li, 2002; Daille &

[1] Note that bilingual terminology alignment has a narrower focus than *bilingual terminology extraction*, but the two terms are often used interchangeably in various papers. The latter covers extraction and alignment of terms between languages.
[2] However, some publicly available parallel corpora do exist. A good overview can be found at the OPUS web portal (Tiedemann, 2012).

Morin, 2005; Morin et al., 2008; Vintar, 2010; Bouamor et al., 2013; Hazem & Morin, 2016, 2017).

The approach designed by Aker et al. (2013) and replicated and adapted in Repar et al. (2019) served as the basis of our work. It was developed to align terminology between languages with the help of parallel corpora using machine-learning techniques. They use terms from the Eurovoc (Steinberger et al., 2002) thesaurus and train an SVM binary classifier (Joachims, 2002) (with a linear kernel and the trade-off between training error and margin parameter c = 10). The task of bilingual alignment is treated as a binary classification - each term from the source language S is paired with each term from the target language T and the classifier then decides whether the aligned pair is correct or incorrect. Aker et al. (2013) run their experiments on the 21 official EU languages covered by Eurovoc with English always being the source language (20 language pairs altogether). They evaluate the performance on a held-out term pair list from Eurovoc using recall, precision and F-measure for all 21 languages. Next, they propose an experimental setting for a simulation of a real-world scenario where they collect English-German comparable corpora of two domains (IT, automotive) from Wikipedia, perform monolingual term extraction using the system by Pinnis et al. (2012) followed by the bilingual alignment procedure described above and manually evaluate the results (using two evaluators). They report excellent performance on the held-out term list with many language pairs reaching 100% precision and the lowest recall being 65%. For Slovenian, which is of our main interest, the reported results were excellent with perfect or nearly perfect precision and good recall. The reported results of the manual evaluation phase were also good, with two evaluators agreeing that at least 81% of the extracted term pairs in the IT domain and at least 60% of the extracted term pairs in the automotive domain can be considered exact translations. Repar et al. (2019) tried to reproduce their approach and after initially having little success they were at the end able to achieve comparable results with precision exceeding 90% and recall over 50%.

Despite the problem of bilingual term alignment lending itself well to the binary classification task, there have been relatively few approaches utilizing machine learning. Similar to Aker et al. (2013), Baldwin & Tanaka (2004) generate corpus-based, dictionary-based and translation-based features and train an SVM classifier to rank the translation candidates. Note that they only focus on multi-word noun phrases (noun + noun). A similar approach, again focusing on noun phrases, is also described by Cao & Li (2002). Finally, Nassirudin & Purwarianti (2015) also reimplement Aker et al. (2013) for the Indonesian-Japanese language pair and further expand it with additional statistical features.

This paper is organized as follows: the present section introduces the problem and related work, Section 2 describes the datasets used for the experiments, Section 3 lists the features used in the machine learning process, Section 4 contains a description of the experiments and lists their results and section 5 provides the conclusion.

2. Resources

The approach described in this paper requires four types of resources. The first two are the same as in Aker et al. (2013) and Repar et al. (2019), whereas the third and fourth resources are required for the additional experiments conducted for this paper:

- aligned term pairs in two languages that serve as training data

- a parallel corpus to generate a Giza++ word alignment list
- pretrained embeddings in two languages
- a (small) bilingual dictionary

We create term pairs from the Eurovoc (Steinberger et al., 2002) thesaurus, which at the time of Repar et al. (2019) consisted of 7,083[3] terms, by pairing Slovenian terms with English ones. The test set consisted of 600 positive (correct) term pairs—taken randomly out of the total 7,083 Eurovoc term pairs—and around 1.3 million negative pairs which were created by pairing each source term with 200 distinct incorrect random target terms. Aker et al. (2013) argue that this was done to simulate real-world conditions where the classifier would be faced with a larger number of negative pairs and a comparably small number of positive ones. The 600 positive term pairs were further divided into 200 pairs where both (i.e. source and target) terms were single words, 200 pairs with a single word only on one side and 200 pairs with multiple-word terms on both sides. The remaining positive term pairs (approximately 6,200) were used as training data along with additional 6,200 negative pairs. These were constructed by taking the source side terms and pairing each source term with one target term (other than the correct one). Using Giza++, we created source-to-target and target-to-source word alignment dictionaries based on the DGT translation memory (Steinberger et al., 2013). The resulting dictionary entries consist of the source word s, its translation t and the number indicating the probability that t is an actual translation of s. To improve the performance of the dictionary-based features, the following entries were removed from the dictionaries:

- entries where probability is lower then 0.05
- entries where the source word was less than 4 characters and the target word more than 5 characters long and vice versa in order to avoid translations of stop word to content words)

In addition to the resources described above, we used fastText (Bojanowski et al., 2016) pre-trained word embedding vectors to calculate distances (or similarities) between terms. We aligned monolingual fastText embeddings using the VecMap (Artetxe et al., 2018) tool which can align embeddings with the help of a small bilingual dictionary. We used a bilingual dictionary compiled from two sources: single word terms from Eurovoc and Wiktionary entries extracted using wikt2dict tool (Acs, 2014). Using the aligned embedding vectors, we then calculated cosine distances between all words present in Eurovoc terms in one language and all words present in Eurovoc terms in the other language.

Using the fastText-based lists of aligned words, we created 3-tuples[4] of most similar - based on cosine similarity - source-to-target and target-to-source words, such as:

- ksenofobija ['xenophobia', '0.744'], ['racism', '0.6797'], ['anti-semitism', '0.654']
- ženska ['woman', '0.7896'], ['women', '0.73'], ['female', '0.722']

where the tuple contains the source language word along with their three most likely corresponding words in the target language and their cosine similarities. The 3-tuples of

[3] While new terms are constantly added to Eurovoc, we decided not to use them to allow for better comparison between the approaches

[4] This number was determined experimentally.

most similar words were used to construct additional features for the machine learning algorithm.

3. Feature construction

The updated approach in this paper uses three types of features that express correspondences between the words (composing a term) in the target and source language. The dictionary and cognate-based features are same as in Repar et al. (2019), while embedding-based features are newly developed. The three feature types are (for a detailed description see Table 1):

- 7 dictionary-based (using Giza++) features which take advantage of dictionaries created from large parallel corpora of which 6 are direction-dependent (source-to-target or target-to-source) and 1 direction-independent - resulting in altogether 13 features
- 7 cognate-based features (on the basis of Gaizauskas et al. (2012)) which utilize string-based word similarity between languages
- 5 cognate-based features using specific transliteration rules which take into account the differences in writing systems between two languages: e.g. Slovenian and English. Transliteration rules were created for both directions (source-to-target and target-to-source) separately and cognate-based features were constructed for both directions - resulting in additional 10 cognate-based features with transliteration rules. The following transliteration rules were used: *x:ks, y:j, w:v, q:k* for English to Slovenian and *č:ch, š:sh, ž:zh* for Slovenian to English
- 5 direction-dependent combined[5] features where the term pair alignment is correct if either the dictionary or the cognate-based method returns a positive result - resulting in a total of 10 combined features
- 12 novel direction-dependent embedding-based features utilizing fastText embeddings - resulting in a total of 24 features
- 5 novel combined features constructed in the same manner as the exisiting combined features but replacing Giza++ word lists with fastText-based lists of top 3 aligned words - resulting in a total of 10 novel combined features
- 3 term length features: sourceTargetLengthMatch, sourceTermLength, targetTermLength

To match words with morphological differences, we do not perform direct string matching but utilize Levenshtein Distance. Two words were considered equal if the Levenshtein Distance Levenshtein (1966) was equal or higher than 0.95.

[5] For combined features, a word is considered as covered if it can be found in the corresponding set of Giza++ translations or if one of the cognate-based measures (Longest Common Subsequence, Longest Common Substring, Levensthein Distance, Needleman-Wunsch Distance, Dice) is 0.70 or higher (set experimentally by Aker et al. (2013))

Feature	Cat	Description
isFirstWordTranslated	Dict	Checks whether the first word of the source term is a translation of the first word in the target term (based on the Giza++ dictionary)
isLastWordTranslated	Dict	Checks whether the last word of the source term is a translation of the last word in the target term
percentageOfTranslatedWords	Dict	Ratio of source words that have a translation in the target term
percentageOfNotTranslatedWords	Dict	Ratio of source words that do not have a translation in the target term
longestTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which has a translation in the target term (compared to the source term length)
longestNotTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which do not have a translation in the target term (compared to the source term length)
Longest Common Subsequence Ratio	Cogn	Measures the longest common non-consecutive sequence of characters between two strings
Longest Common Substring Ratio	Cogn	Measures the longest common consecutive string (LCST) of characters that two strings have in common
Dice similarity	Cogn	2*LCST / (len(source) + len(target))
Needlemann-Wunsch distance	Cogn	LCST / min(len(source), len(target))
isFirstWordCognate	Cogn	A binary feature which returns True if the longest common consecutive string (LCST) of the first words in the source and target terms divided by the length of the longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters
isLastWordCognate	Cogn	A binary feature which returns True if the longest common consecutive string (LCST) of the last words in the source and target terms divided by the length of longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters
Normalized Levensthein distance (LD)	Cogn	1 - LD / max(len(source), len(target))
isFirstWordCovered	Comb	A binary feature indicating whether the first word in the source term has a translation or transliteration in the target term
isLastWordCovered	Comb	A binary feature indicating whether the last word in the source term has a translation or transliteration in the target term
percentageOfCoverage	Comb	Returns the percentage of source term words which have a translation or transliteration in the target term
percentageOfNonCoverage	Comb	Returns the percentage of source term words which have neither a translation nor transliteration in the target term
difBetweenCoverageAndNonCoverage	Comb	Returns the difference between the last two features
isFirstWordMatch	Emd	Checks whether the first word of the source term is the most likely translation of the first word in the target term (based on the aligned embeddings)
isLastWordMatch	Emd	Checks whether the last word of the source term is the most likely translation of the last word in the target term (based on the aligned embeddings)
percentageOfFirstMatchWords	Emb	Ratio of source words that have a first match (i.e. first position in the 3-tuple) in the target term
percentageOfNotFirstMatchWords	Emb	Ratio of source words that do not have a first match (i.e. first position in the 3-tuple) in the target term
longestFirstMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which has a first match (first position in the 3-tuple) in the target term (compared to the source term length)
longestNotFirstMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which do not have a first match (first position in the 3-tuple) in the target term (compared to the source term length)
isFirstWordTopnMatch	Emd	Checks whether the first word of the source term is in the 3-tuple of most likely translations of the first word in the target term (based on the aligned embeddings)

isLastWordTopnMatch	Emd	Checks whether the first word of the source term is not in the 3-tuple of most likely translations of the first word in the target term (based on the aligned embeddings)
percentageOfTopnMatchWords	Emb	Ratio of source words that have a match (i.e. any position in the 3-tuple) in the target term
percentageOfNotTopnMatchWords	Emb	Ratio of source words that do not have a match (i.e. any position in the 3-tuple) in the target term
longestTopnMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which has a match (any position in the 3-tuple) in the target term (compared to the source term length)
longestNotTopnMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which do not have a match (any position in the 3-tuple) in the target term (compared to the source term length)
isFirstWordCoveredEmbeddings	Comb	A binary feature indicating whether the first word in the source term has a match (any position in the 3-tuple) or transliteration in the target term
isLastWordCoveredEmbeddings	Comb	A binary feature indicating whether the last word in the source term has a match (any position in the 3-tuple) or transliteration in the target term
percentageOfCoverageEmbeddings	Comb	Returns the percentage of source term words which have a match (any position in the 3-tuple) or transliteration in the target term
percentageOfNonCoverageEmbeddings	Comb	Returns the percentage of source term words which do not have a match (any position in the 3-tuple) or transliteration in the target term
diffBetweenCoverageAnd-NonCoverageEmbeddings	Comb	Returns the difference between the last two features

Figure 1: Features used in the experiments. Note that some features are used more than once because they are direction-dependent.

4. Experimental setup and results

The constructed features were then used to train an SVM binary classifier (Joachims, 2002) (with a linear kernel and the trade-off between training error and margin parameter c = 10). We selected three configurations from Repar et al. (2019) for comparison:

- Training set 1:200: a very unbalanced train set (ratio of 1:200 between positive and negative examples [6]) greatly improves the precision of the classifier at a cost of somewhat lower recall, when compared to a balanced train set or a less unbalanced train set (e.g., ratio of 1:10 between positive and negative examples).
- Training set filtering 3: In Repar et al. (2019), we have performed an error analysis and found that many incorrectly classified term pairs are cases of partial translation where one unit in a multi-word term has a correct Giza++ dictionary translation in the corresponding term in the other language. Based on the problem of partial translations, leading to false positive examples, we focused on the features that would eliminate such partial translations from the training set. After a systematic experimentation, we noticed that we can drastically improve precision if we only keep positive term pairs with the following feature values: isFirstWordTranslated = True, isLastWordTranslated = True, percentageOfCoverage > 0.66, isFirstWordTranslated-reversed = True, isLastWordTranslated-reversed = True, percentageOfCoverage-reversed > 0.66.

[6] 1:200 imbalance ratio was the largest imbalance we tried, since the testing results indicated that no further gains could be achieved by further increasing the imbalance.

- Cognates: the dataset is additionally filtered according to the following criteria: isFirstWordCognate = True and isLastWordCognate = True, isFirstWordTranslated = True and isLastWordCognate = True, isFirstWordCognate = True and isLastWordTranslated = True and we also use a Gaussian kernel instead of the linear one, since this new dataset structure represents a classic "exclusive or" (XOR) problem which a linear classifier is unable to solve.

The selection was made based on our experience and previous work with this approach. The three selected configurations were among the best performing in previous experiments and we believed they had the highest potential for improvement. For a complete description of the decisions that led to these configurations, please refer to Repar et al. (2019).

No.	Config EN-SL	Training set size	Pos/Neg ratio	Precision	Recall	F-score
Dictionary-based and cognate-based features						
1	Training set 1:200	1,303,083	1:200	0.4299	**0.7617**	0.5496
2	Training set filtering 3	645,813	1:200	0.9342	0.4966	0.6485
3	Cognates approach	672,345	1:200	0.8732	0.5167	0.6492
Dictionary-based, embedding-based and cognate-based features						
1	Training set 1:200	1,303,083	1:200	0.5375	0.680	0.6004
2	Training set filtering 3	695,058	1:200	0.8170	0.5133	0.6305
3	Cognates approach	706,113	1:200	0.8991	0.5200	**0.6589**
Embedding-based and cognate-based features only						
1	Training set 1:200	1,303,083	1:200	0.3232	0.4967	0.3916
2	Training set filtering 3	322,605	1:200	0.9545	0.2450	0.3899
3	Cognates approach	394,362	1:200	**0.9618**	0.3617	0.5242

Table 2: Results on the English-Slovenian term pair.

First, we simply added the new embedding-based features to the dataset to see if they improved the overall performance. Later, we removed the dictionary-based features from the dataset to see whether the novel embedding-based features could replace them without a major impact to the performance. As can be observed from Table 2, the results are a mixed bag when using all available features. Without any training set filtering, the new features improve precision at the expense of recall, but are less effective when filtering is applied. Nevertheless, when we use additional trainset filters for the Cognates approach, we can observe a slight increase in both precision and recall resulting in the overall highest F-score. When we use only embedding-based and cognate-based features, which would be beneficial for language pairs without access to large parallel corpora needed to create Giza++ word alignments, there is a significant drop in recall in all cases, but precision actually increases when trainset filtering is applied and the Cognates approach achieves the overall best precision.

5. Conclusion

In this paper, we continued our experiments on bilingual terminology alignment using a machine learning approach by adding new features based on fastText word embedding

vectors. We took advantage of the availability of large pre-trained datasets by Bojanowski et al. (2016), and a cross-lingual word embedding mapping tool Vecmap by Artetxe et al. (2018) to create word alignment dictionaries similar to the output of traditional word alignment tools, such as Giza++ (Och & Ney, 2003). The single most important advantage of this approach is that while Giza++ requires a large parallel corpus, fastText vectors are trained on monolingual data and Vecmap needs only a (much smaller) bilingual dictionary. Bilingual dictionaries are readily available for many language pairs via Wiktionary (Acs, 2014).

The experiments showed that the new features can have a positive impact on F-score (depending on the configuration), but precision was somewhat lower compared to when we were using only Giza++ features. When we removed Giza++ features and using only the new embedding-based features (alongside cognate features which are based on word similarity and require no pre-existing bilingual data), we observed somewhat lower recall and a bit higher precision. This means that the embedding-based features can be used instead of Giza++ features for language pairs where no large parallel bilingual corpora is available.

In terms of future work, we plan on creating additional features using contextual embeddings, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), which could potential help us improve recall, and explore more granular and detailed training set filtering techniques. We also plan to expand the experiments and test other configurations in a more systematic way.

6. Acknowledgements

The work was supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). We also acknowledge the project the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103) and the project TermFrame - Terminology and Knowledge Frames across Languages (No. J6-9372).

7. References

Acs, J. (2014). Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC*.

Aker, A., Paramita, M. & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1. pp. 402–411.

Artetxe, M., Labaka, G. & Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Baldwin, T. & Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. pp. 24–31.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Bouamor, D., Semmar, N. & Zweigenbaum, P. (2013). Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pp. 759–764.

Cao, Y. & Li, H. (2002). Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*. pp. 1–7.

Chiao, Y.C. & Zweigenbaum, P. (2002). Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2*. pp. 1–5.

Daille, B., Gaussier, E. & Langé, J.M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*. pp. 515–521.

Daille, B. & Morin, E. (2005). French-English Terminology Extraction from Comparable Corpora. In *Natural Language Processing – IJCNLP 2005*. pp. 707–718.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Fung, P. & Yee, L.Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. pp. 414–420.

Gaizauskas, R., Aker, A. & Yang Feng, R. (2012). Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics*. pp. 23–32.

Gaussier, E. (1998). Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. pp. 444–450.

Hazem, A. & Morin, E. (2016). Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pp. 3401–3411.

Hazem, A. & Morin, E. (2017). Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. pp. 685–693.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. pp. 17–22.

Levenshtein, V.I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, p. 707.

Morin, E., Daille, B., Takeuchi, K. & Kageura, K. (2008). Brains, Not Brawn: The Use of Smart Comparable Corpora in Bilingual Terminology Mining. *ACM Trans. Speech Lang. Process.*, 7(1), pp. 1:1–1:23.

Nassirudin, M. & Purwarianti, A. (2015). Indonesian-Japanese term extraction from bilingual corpora using machine learning. In *Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on*. pp. 111–116.

Och, F.J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), pp. 19–51.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Pinnis, M., Ljubešić, N., Stefanescu, D., Skadina, I., Tadić, M. & Gornostaya, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*. pp. 20–21.

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. pp. 519–526.

Repar, A., Martinc, M. & Pollak, S. (2019). Reproduction, replication, analysis and adaptation of a term alignment approach. *Language Resources and Evaluation*, pp. 1–34.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S. & Schlüter, P. (2013). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*.

Steinberger, R., Pouliquen, B. & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, pp. 101–121.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N.C.C. Chair), K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Vintar, Š. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(2), pp. 141–158.

3.4 Classification with Sentence Embeddings

3.4.1 Description of the Approach

Finally, in Repar et al. (2022), we extended the approach by adding sentence-transformer features generated with the methodology described in Reimers and Gurevych (2019b). We consider using the sentence-transformers because of the sub-word information that is taken into account while learning the model. In addition to the dictionary, cognate and embedding-based features, we feed the model with single or multi-word terms as "sentences" and obtain the sentence-embedding features. For each term in each language respectively we obtain the sentence embeddings and then for each term in English we rank all of the French terms with regards of cosine similarity.

3.4.2 Results

Using just sentence embeddings in isolation (average precision of 0.680) did not improve on the results of the approach described in Repar et al. (2021) (average precision of 0.712), but combining both the old approach and the new sentence-embedding features in a machine learning setting did result in improved performance (average precision of 0.833). Note that this paper used a different evaluation methodology because it was prepared for the BUCC Shared Task.

Fusion of linguistic, neural and sentence-transformer features for improved term alignment

Andraž Repar[1], **Boshko Koloski**[1], **Matej Ulčar**[2], **Senja Pollak**[1]
[1]Jožef Stefan Institute, Jožef Stefan International Postgraduate School
Jamova cesta 39, Ljubljana, Slovenia
[2]Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
{andraz.repar,boshko.koloski,senja.pollak}@ijs.si, matej.ulcar@fri.uni-lj.si

Abstract

Crosslingual terminology alignment task has many practical applications. In this work, we propose an aligning method for the shared task of the 15th Workshop on Building and Using Comparable Corpora. Our method combines several different approaches into one cohesive machine learning model, based on SVM. From shared-task specific and external sources, we crafted four types of features: cognate-based, dictionary-based, embedding-based, and combined features, which combine aspects of the other three types. We added a post-processing re-scoring method, which reduces the effect of hubness, where some terms are nearest neighbours of many other terms. We achieved the average precision score of 0.833 on the English-French training set of the shared task.

Keywords: term alignment, cognates, embeddings, sentence-transformers

1. Introduction

Having the ability to align concepts between languages can provide significant benefits in many practical applications, such as aligning terms between languages in bilingual terminology, aligning keywords in news industry or using aligned concepts as seed data for other NLP tasks like multilingual vector space alignment.

In this paper, we present the experiments and their results on the data provided in the bilingual term alignment in comparable specialized corpora shared task organized as part of the 15th Workshop on Building and Using Comparable Corpora (the BUCC workshop). Given a pair of comparable corpora in two languages and a pair of term lists where terms originate in the two corpora, participants were required to produce lists of term pair candidates ranked by their alignment probability (i.e. terms closer to the top are more likely to be true alignments).

Our method involves a machine learning approach based on our work in (Repar et al., 2019) and (Repar et al., 2021) with additional improvements. Our system uses several external resources detailed in Section 3, all of which are publicly available online.

This paper is organized as follows: Section 1 introduces the topic, Section 2 provides the related work, Section 3 describes the methodology, Section 4 contains the results and Section 5 the conclusion.

2. Related work

Initial attempts at bilingual terminology extraction involved parallel input data (Kupiec, 1993; Daille et al., 1994; Gaussier, 1998), and the interest of the community continued until today. However, most parallel corpora are owned by private companies[1], such as language service providers, who consider them to be their intellectual property and are reluctant to share them publicly. For this reason (and in particular for language pairs not involving English) considerable efforts have also been invested into researching bilingual terminology extraction from comparable corpora (Fung and Yee, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Cao and Li, 2002; Daille and Morin, 2005; Morin et al., 2008; Vintar, 2010; Bouamor et al., 2013a; Bouamor et al., 2013b; Hazem and Morin, 2016; Hazem and Morin, 2017).

The approach designed by Aker et al. (2013) and replicated and adapted in Repar et al. (2019) served as the basis of our work. It was developed to align terminology between languages with the help of parallel corpora using machine-learning techniques. They use terms from the Eurovoc (Steinberger et al., 2002) thesaurus and train an SVM binary classifier (Joachims, 2002) (with a linear kernel and the trade-off between training error and margin parameter c = 10). The task of bilingual alignment is treated as a binary classification - each term from the source language *S* is paired with each term from the target language *T* and the classifier then decides whether the aligned pair is correct or incorrect. Aker et al. (2013) run their experiments on the 21 official EU languages covered by Eurovoc with English always being the source language (20 language pairs altogether). They evaluate the performance on a held-out term pair list from Eurovoc using recall, precision and F-measure for all 21 languages. Next, they

[1]However, some publicly available parallel corpora do exist. A good overview can be found at the OPUS web portal (Tiedemann, 2012).

propose an experimental setting for a simulation of a real-world scenario where they collect English-German comparable corpora of two domains (IT, automotive) from Wikipedia, perform monolingual term extraction using the system by Pinnis et al. (2012) followed by the bilingual alignment procedure described above and manually evaluate the results (using two evaluators). They report excellent performance on the held-out term list with many language pairs reaching 100% precision and the lowest recall being 65%. For Slovenian, which is of our main interest, the reported results were excellent with perfect or nearly perfect precision and good recall. The reported results of the manual evaluation phase were also good, with two evaluators agreeing that at least 81% of the extracted term pairs in the IT domain and at least 60% of the extracted term pairs in the automotive domain can be considered exact translations. Repar et al. (2019) tried to reproduce their approach and after initially having little success they were at the end able to achieve comparable results with precision exceeding 90% and recall over 50%.

Despite the problem of bilingual term alignment lending itself well to the binary classification task, there have been relatively few approaches utilizing machine learning. Similar to Aker et al. (2013), Baldwin and Tanaka (2004) generate corpus-based, dictionary-based and translation-based features and train an SVM classifier to rank the translation candidates. Note that they only focus on multi-word noun phrases (noun + noun). A similar approach, again focusing on noun phrases, is also described by Cao and Li (2002). Finally, Nassirudin and Purwarianti (2015) also reimplement Aker et al. (2013) for the Indonesian-Japanese language pair and further expand it with additional statistical features.

3. Methodology

Initial experiments were performed with cross-lingual embeddings (see Section 3.1) and sentence transformers (see Section 3.2). However, the results were lower than expected, which is why we adapted an approach described in Repar et al. (2021) by adding additional features based on the cross-lingual embedding and sentence transformer experiments.

3.1. Cross-lingual aligned embeddings

We used fastText Bojanowski et al. (2017) word embeddings for both involved languages. We constructed a bilingual English-French dictionary from Wiktionary entries, using the wikt2dict tool Acs (2014). The extracted dictionary has 204 341 entries. For the purpose of embedding alignment, we filtered it to keep only single-word entries, i.e. those that have a single word in both languages. After the filtering, we had 129 912 entries, of which 24 923 have an identical word in both languages (e.g. place names or chemicals) There's an average of 1.55 English translations for each French word, and 1.56 French translations for each English word. 23.4% of English words have multiple French translations, while 24.3% of French words have multiple English translations.

We then aligned the French and English word embeddings into a common vector space in a supervised manner, utilizing the bilingual dictionary. We used Vecmap Artetxe et al. (2018) tool, which aligns the vectors using the Moore-Penrose pseudo-inverse, which minimizes the sum of squared Euclidean distances. We extracted one vector for each term in each language. For multi-word terms we averaged the word vectors of all the words the term is composed of. Finally we use the cosine similarity score to find the most similar terms in language 1 for each term in language 2, and vice-versa. Using this approach, we achieve an average precision of 0.496 (for details, see Table 2).

3.2. Sentence-transformers features

We used the Sentence-Transformers Reimers and Gurevych (2019) model to embed the terms of the both languages. We utilized the implementations of *c19 python library* (Koloski et al., 2021) to obtain the embeddings [2]. The sentence-transformer architecture is designed to solve the task of sentence similarity, it leverages BERT tokens and via pooling it creates sentence-embeddings. The BERT Devlin et al. (2018) model uses tokens as input to it's transformer architecture, the BERT-tokenizer tokenizes the words in subwords. We consider using the sentence-transformers because of the sub-word information that is taken into account while learning the model. We feed the model with single or multi-word terms as "sentences" and obtain the sentence-embedding.

3.2.1. Terms as sentences evaluation methodology

For each term in each language respectively we obtain the sentence-embeddings. Next, for each term in English we rank all of the French terms with regards of cosine-similarity.

We consider using five different Language Models:

- XLM (Lample and Conneau, 2019)
- DistilBERT (Sanh et al., 2019)
- All-MPNet (Song et al., 2020)
- MiniLM (Wang et al., 2020)
- Roberta-Large (Liu et al., 2019)

The highest average precision of 0.680 among the five models was achieved with the *distilbert-base* model (for details, see Table 2).

3.3. Supervised machine learning approach

Since the results of the individual approaches described in the previous two sections were lower than expected, we further experimented with combining the individual models into a machine learning model. We reused and

[2] https://github.com/bkolosk1/c19_rep

adapted an approach described in Repar et al. (2021) by incorporating the cosine similarity values of the cross-lingual and sentence transformer models into features of the machine learning model.

This approach uses Eurovoc (Steinberger et al., 2002) terms, Giza++ dictionaries (generated from the DGT translation memory (Steinberger et al., 2013)) and word similarity information to generate features for an SVM binary classifier (Joachims, 2002) (with the trade-off between training error and margin parameter c = 10). The model is trained on a list of 7181 Eurovoc English-French term pairs as well as an additional 1.4 million incorrect term pairs generated by randomly pairing English and French Eurovoc terms to simulate real-world conditions. In addition to the binary classification, the model also provides confidence scores which are later used to rank aligned candidate pairs.

For each potential candidate pair, we calculate features of the following types:

- Cognate-based features
- Dictionary-based features
- Embedding-based features
- Combined features

As described in Repar et al. (2019) and Repar et al. (2021), cognate-based features take advantage of word similarity between languages (e.g. *democracy* in English and *démocratie* in French) and dictionary-based features are calculated using results of the Giza++ word alignment algorithm. Embedding-based features are calculated using cosine similarity scores described in Sections 3.1 and 3.2. For each model, we produce a list of word pairs with their cosine similarity scores. These scores are then used to generate embedding features by creating 3-tuples[3] of most similar - based on cosine similarity - source-to-target and target-to-source words, such as:

- xénophobie ['xenophobia', '0.744'], ['racism', '0.6797'], ['anti-semitism', '0.654']
- femme ['woman', '0.7896'], ['women', '0.73'], ['female', '0.722']

where the tuple contains the source language word along with their three most likely corresponding words in the target language and their cosine similarities. The 3-tuples of most similar words were used to construct additional features for the machine learning algorithm as indicated in 1. Finally, combined features combine some aspects of the first three feature types.

[3]This number was determined experimentally.

3.4. Post-process re-ordering

In post-processing we altered the confidence scores of some of the term-pairings. For some term x_1 from language 1, we wanted to ensure that the best performing aligned pair is as close to the top of the list as possible. For x_1, a large number of candidate terms from language 2 can have a high confidence score for a matching term and this might negatively affect the final average precision scores as defined in the shared task, since most terms would not have more than 2-3 correct alignments. Another term x_2 from language 1 might have a lower confidence score with every candidate term from language 2 than all the candidates for x_1. That is, there are such $x_1, x_2 \in L_1$, that $S(x_1, y) > \max_{y'}(S(x_2, y')), \forall y \in L_2$, where S is confidence/similarity score and L_1 and L_2 are languages 1 and 2, respectively. We therefore boosted the confidence scores of the top n candidates for each term by a constant c. Based on the performance on the training dataset, we chose the parameters $n = 1$ and $c = 1.0$.

4. Experimental setup

In step one, we trained the model on publicly available data (Eurovoc thesaurus, Giza++ word alignment lists trained on the DGT corpus and embedding and transformer models trained on the data provided within the BUCC shared task). In step two, we evaluated its performance on the term lists provided as part of the training package in the shared task. To do so, we generated all possible term pairs between the English and French term lists, calculated the features described in Table 1, produced predictions using the model trained in step one and evaluated them against the English-French term list provided as part of the shared task training data.

5. Results

We report results in Table 2. Using just individual language models described in Section 3.2, the best average precision (0.680) is achieved with the distilbert-base model. When we used the SVM approach described in Repar et al. (2021) (i.e. the *SVM old*, we reach an average precision of 0.712 and when we add additional features based on sentence transformer models we achieve an average precision of 0.833 (i.e. *SVM new*. The post-process re-ordering parameters *n* and *c* were as indicated in Section 3.4.

6. Conclusion

In this paper, we presented the results of our experiments for the shared task of the 15th Workshop on Building and Using Comparable Corpora. We first attempted to align terms using cross-lingual embedding and sentence transformer models, but the results were less than satisfactory. Next, we reused an existing machine learning approach and added additional features based on the cross-lingual embedding and sentence

Feature	Cat	Description
isFirstWordTranslated	Dict	Checks whether the first word of the source term is a translation of the first word in the target term (based on the Giza++ dictionary)
isLastWordTranslated	Dict	Checks whether the last word of the source term is a translation of the last word in the target term
percentageOfTranslatedWords	Dict	Ratio of source words that have a translation in the target term
percentageOfNotTranslatedWords	Dict	Ratio of source words that do not have a translation in the target term
longestTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which has a translation in the target term (compared to the source term length)
longestNotTranslatedUnitInPercentage	Dict	Ratio of the longest contiguous sequence of source words which do not have a translation in the target term (compared to the source term length)
Longest Common Subsequence Ratio	Cogn	Measures the longest common non-consecutive sequence of characters between two strings
Longest Common Substring Ratio	Cogn	Measures the longest common consecutive string (LCST) of characters that two strings have in common
Dice similarity	Cogn	2*LCST / (len(source) + len(target))
Needlemann-Wunsch distance	Cogn	LCST / min(len(source), len(target))
isFirstWordCognate	Cogn	A binary feature which returns True if the longest common consecutive string (LCST) of the first words in the source and target terms divided by the length of the longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters
isLastWordCognate	Cogn	A binary feature which returns True if the longest common consecutive string (LCST) of the last words in the source and target terms divided by the length of longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters
Normalized Levensthein distance (LD)	Cogn	1 - LD / max(len(source), len(target))
isFirstWordCovered	Comb	A binary feature indicating whether the first word in the source term has a translation or transliteration in the target term
isLastWordCovered	Comb	A binary feature indicating whether the last word in the source term has a translation or transliteration in the target term
percentageOfCoverage	Comb	Returns the percentage of source term words which have a translation or transliteration in the target term
percentageOfNonCoverage	Comb	Returns the percentage of source term words which have neither a translation nor transliteration in the target term
difBetweenCoverageAndNonCoverage	Comb	Returns the difference between the last two features
isFirstWordMatch	Emd	Checks whether the first word of the source term is the most likely translation of the first word in the target term (based on the aligned embeddings)
isLastWordMatch	Emd	Checks whether the last word of the source term is the most likely translation of the last word in the target term (based on the aligned embeddings)
percentageOfFirstMatchWords	Emb	Ratio of source words that have a first match (i.e. first position in the 3-tuple) in the target term
percentageOfNotFirstMatchWords	Emb	Ratio of source words that do not have a first match (i.e. first position in the 3-tuple) in the target term
longestFirstMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which has a first match (first position in the 3-tuple) in the target term (compared to the source term length)
longestNotFirstMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which do not have a first match (first position in the 3-tuple) in the target term (compared to the source term length)
isFirstWordTopnMatch	Emd	Checks whether the first word of the source term is in the 3-tuple of most likely translations of the first word in the target term (based on the aligned embeddings)
isLastWordTopnMatch	Emd	Checks whether the first word of the source term is not in the 3-tuple of most likely translations of the first word in the target term (based on the aligned embeddings)
percentageOfTopnMatchWords	Emb	Ratio of source words that have a match (i.e. any position in the 3-tuple) in the target term
percentageOfNotTopnMatchWords	Emb	Ratio of source words that do not have a match (i.e. any position in the 3-tuple) in the target term
longestTopnMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which has a match (any position in the 3-tuple) in the target term (compared to the source term length)
longestNotTopnMatchUnitInPercentage	Emb	Ratio of the longest contiguous sequence of source words which do not have a match (any position in the 3-tuple) in the target term (compared to the source term length)
isFirstWordCoveredEmbeddings	Comb	A binary feature indicating whether the first word in the source term has a match (any position in the 3-tuple) or transliteration in the target term
isLastWordCoveredEmbeddings	Comb	A binary feature indicating whether the last word in the source term has a match (any position in the 3-tuple) or transliteration in the target term
percentageOfCoverageEmbeddings	Comb	Returns the percentage of source term words which have a match (any position in the 3-tuple) or transliteration in the target term
percentageOfNonCoverageEmbeddings	Comb	Returns the percentage of source term words which do not have a match (any position in the 3-tuple) or transliteration in the target term
diffBetweenCoverageAnd-NonCoverageEmbeddings	Comb	Returns the difference between the last two features

Table 1: Features used in the experiments. Note that some features are used more than once because they are direction-dependent or used multiple times with different embedding or transformer models.

Model	Average precision
aligned fastText	0.496
distilbert-base	0.680
xlm-r	0.650
all-mpnet	0.616
all-MiniLM	0.621
roberta-large	0.523
SVM old	0.712
SVM new	**0.833**

Table 2: Results

transformer models. Using this model, we achieved the average precision of 0.833. Our experiments show that careful feature engineering could still produce better results than more novel deep learning approaches.

In terms of future work, there is still room for improvement which could be achieved by generating additional features using other transformer or embedding models. The system is also quite resource intensive — model training and prediction on the BUCC dataset took more than 24 hours. Finally, there is also room for a more systematic approach to the postprocess re-ranking step.

7. Acknowledgements

This work was supported by the Slovenian Research Agency (ARRS) grants for the core programme Knowledge technologies (P2-0103), as well as the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

8. Bibliographical References

Acs, J. (2014). Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC*.

Aker, A., Paramita, M., and Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 402–411.

Artetxe, M., Labaka, G., and Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Baldwin, T. and Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In

Proceedings of the Workshop on Multiword Expressions: Integrating Processing, pages 24–31.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bouamor, D., Popescu, A., Semmar, N., and Zweigenbaum, P. (2013a). Building specialized bilingual lexicons using large scale background knowledge. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 479–489, Seattle, Washington, USA, October. Association for Computational Linguistics.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2013b). Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 759–764.

Cao, Y. and Li, H. (2002). Base noun phrase translation using web data and the EM algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, pages 1–7.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2*, pages 1–5.

Daille, B. and Morin, E. (2005). French-English terminology extraction from comparable corpora. In *Natural Language Processing – IJCNLP 2005*, pages 707–718.

Daille, B., Gaussier, E., and Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, pages 515–521.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, pages 414–420.

Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, pages 444–450.

Hazem, A. and Morin, E. (2016). Efficient data selection for bilingual terminology extraction from comparable corpora. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3401–3411.

Hazem, A. and Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 685–693.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

Koloski, B., Stepišnik-Perdih, T., Pollak, S., and Škrlj, B. (2021). Identification of covid-19 related fake news via neural stacking. In Tanmoy Chakraborty, et al., editors, *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 177–188, Cham. Springer International Publishing.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 17–22.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2008). Brains, not brawn: The use of smart comparable corpora in bilingual terminology mining. *ACM Trans. Speech Lang. Process.*, 7(1):1:1–1:23, October.

Nassirudin, M. and Purwarianti, A. (2015). Indonesian-Japanese term extraction from bilingual corpora using machine learning. In *Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on*, pages 111–116.

Pinnis, M., Ljubešić, N., Stefanescu, D., Skadina, I., Tadić, M., and Gornostaya, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, pages 20–21.

Rapp, R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Repar, A., Martinc, M., and Pollak, S. (2019). Reproduction, replication, analysis and adaptation of a term alignment approach. *Language Resources and Evaluation*, pages 1–34.

Repar, A., Martinc, M., Ulčar, M., and Pollak, S. (2021). Word-embedding based bilingual terminology alignment. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, page 98.

Sanh, V., Debut, L., Chaumond, J., and Wolf,

T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. (2020). Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297.

Steinberger, R., Pouliquen, B., and Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, pages 101–121.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2013). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Vintar, Š. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(2):141–158.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

9. Language Resource References

3.5 Terminology Alignment in the Media Industry

3.5.1 Description of the Approach

In this paper, we applied the bilingual terminology alignment methodology described in Section 3.2 to a dataset of unaligned Estonian and Russian keywords which were manually assigned by journalists to describe the article topic. The dataset of Estonian and Russian tags was provided by Ekspress Meedia as a simple list of one tag per line. The total number of tags was 65,830. We made several adjustments to the methodology from Section 3.2:

- We generated special transliteration rules between Latin and Cyrillic script
- For Giza++ training, since Russian is not an EU language, we switched from the DGT translation memory to the Russian-Estonian OpenSubtitles corpus from the Opus portal[2].
- For the training set, since Eurovoc does not contain Russian, we used the environmental corpus Gemet[3]

3.5.2 Results

The domains of the language-specific resources (subtitles for the Giza++ training corpus and environment for the machine learning training dataset) did not match the domain of the dataset (news articles), which results in worse performance compared to the approach in Section 3.2, particularly in terms of recall. Nevertheless, we then used the best performing configuration to try to align the Estonian and Russian tags from the dataset provided by Ekspress Meedia which resulted in 4989 positively classified Estonian-Russian tags. A subset of these (500) were manually evaluated by a person with knowledge of both languages. Of the 500 positively classified tag pairs that were manually evaluated, 49% percent were deemed to be complete matches, a further 25% were evaluated as partial matches, and 26% were considered to be wrongly classified as positive tag pairs.

3.5.3 Relevance of the developed approaches

Aligning terms from specialized corpora across languages is a well-established challenge in the translation industry, which the approaches described above aim to address. However, the findings of Adjali, Morin, Sharoff, et al., 2022 indicate that there remains significant room for improvement. This suggests that methods not originally designed for this specific task—such as neural machine translation and in particular LLM-based translation—may offer valuable contributions and potential advancements in this area.

[2] opus.nlpl.eu

[3] https://www.eionet.europa.eu/gemet/en/themes/

Aligning Estonian and Russian news industry keywords with the help of subtitle translations and an environmental thesaurus

Andraž Repar
International Postgraduate School / Jamova 39, 1000 Ljubljana, Slovenia
andraz.repar@ijs.si
Andrej Shumakov
Ekspress Meedia / Narva mnt 13, 10151 Tallinn, Estonia

Abstract

This paper presents the implementation of a bilingual term alignment approach developed by Repar et al. (2019) to a dataset of unaligned Estonian and Russian keywords which were manually assigned by journalists to describe the article topic. We started by separating the dataset into Estonian and Russian tags based on whether they are written in the Latin or Cyrillic script. Then we selected the available language-specific resources necessary for the alignment system to work. Despite the domains of the language-specific resources (subtitles and environment) not matching the domain of the dataset (news articles), we were able to achieve respectable results with manual evaluation indicating that almost 3/4 of the aligned keyword pairs are at least partial matches.

1 Introduction and related work

The ability to accurately align concepts between languages can provide significant benefits in many practical applications. For example, in terminology, terms can be aligned between languages to provide bilingual terminological resources, while in the news industry, keywords can be aligned to provide better news clustering or search in another language. Accurate bilingual resources can also serve as seed data for various other NLP tasks, such as multilingual vector space alignment.

In this paper, we describe the experiments on an Estonian-Russian dataset of news tags — labels that were manually assigned to news articles by journalists and editors at Ekspress Meedia, one of the largest news publishers in the Baltic region. The dataset contains both Estonian and Russian tags, but they are not aligned between the two languages. We adapted the machine learning term alignment approach described by Repar et al. (2019) to align the Russian and Estonian tags in the dataset.

The alignment approach in Repar et al. (2019) is a reproduction and adaptation of the approach described by Aker et al. (2013a). Repar et al. (2019) managed to reach a precision of over 0.9 and therefore approach the values presented by Aker et al. (2013a) by tweaking several parameters and developing new machine learning features. They also developed a novel cognate-based approach which could be effective in texts with a high proportion of novel terminology that cannot be detected by relying on dictionary-based features. In this work, we perform the implementation of the proposed method on a novel, Estonian-Russian language pair, and in a novel application of tagset alignment.

Section 1 lists the related work, Section 2 contains a description of the tag dataset used, Section 3 describes the system architecture, Section 4 explains the resources used in this paper, Section 5 contains the results of the experiments and Section 6 provides conclusions and future work.

2 Dataset description

The dataset of Estonian and Russian tags was provided by Ekspress Meedia as a simple list of one tag per line. The total number of tags was 65,830. The tagset consists of keywords that journalists assigne to articles to describe an article's topic, and was cut down recently by the editors from more than 210,000 tags.

The number of Russian tags was 6,198 and they were mixed with the Estonian tags in random order. Since Russian and Estonian use different writing scripts (Cyrillic vs Latin), we were able to separate the tags using a simple regular expression to detect Cyrillic characters. The vast majority of the tags are either unigrams or bigrams (see Table 1 for details).

Grams	Estonian	Russian
1	0.49	0.49
2	0.44	0.41
3	0.05	0.06
4	0.01	0.02
> 4	0.01	0.02

Table 1: An analysis of the provided dataset in terms of multi-word units. The values represent the ratio of the total number of tags for the respective language. The total number of Estonian tags was 59,632, and the total number of Russian tags was 6,198. The largest Estonian tag was a 14-gram and the largest Russian tag was an 11-gram, but the vast majority of tags are either unigrams or bigrams.

3 System architecture

The algoritm used in this paper is based on the approach described in Repar et al. (2019) which is itself a replication and an adaptation of Aker et al. (2013b). The original approach designed by (Aker et al., 2013b) was developed to align terminology from comparable (or parallel) corpora using machine-learning techniques. They use terms from the Eurovoc (Steinberger et al., 2002) thesaurus and train an SVM binary classifier (Joachims, 2002) (with a linear kernel and the trade-off between training error and margin parameter c = 10). The task of bilingual alignment is treated as a binary classification - each term from the source language *S* is paired with each term from the target language *T* and the classifier then decides whether the aligned pair is correct or incorrect. (Aker et al., 2013b) use two types of features that express correspondences between the words (composing a term) in the target and source language:

- 7 dictionary-based (using Giza++) features which take advantage of dictionaries created from large parallel corpora of which 6 are direction-dependent (source-to-target or target-to-source) and 1 direction-independent - resulting in altogether 13 features, and
- 5 cognate-based (on the basis of (Gaizauskas et al., 2012)) which utilize string-based word similarity between languages.

To match words with morphological differences, they do not perform direct string matching but utilize Levenshtein Distance. Two words were considered equal if the Levenshtein Distance (Levenshtein, 1966) was equal or higher than 0.95. For closed-compounding languages, they check whether the compound source term has an initial prefix that matches the translation of the first target word, provided that translation is at least 5 characters long.

Additional features are also constructed by:

- Using language pair specific transliteration rules to create additional cognate-based features. The purpose of this task was to try to match the cognate terms while taking into account the differences in writing systems between two languages: e.g. Greek and English. Transliteration rules were created for both directions (source-to-target and target-to-source) separately and cognate-based features were constructed for both directions - resulting in additional 10 cognate-based features with transliteration rules.
- Combining the dictionary and cognate-based features in a set of combined features where the term pair alignment is correct if either the dictionary or the cognate-based method returns a positive result. This process resulted in additional 10 combined features[1].

A subset of the features is described below (For a full list of features, see Repar et al. (2019)):

- *isFirstWordTranslated*: A dictionary feature that checks whether the first word of the source term is a translation of the first word in the target term (based on the Giza++ dictionary).
- *longestTranslatedUnitInPercentage*: A dictionary feature representing the ratio of the longest contiguous sequence of source words which has a translation in the target term (compared to the source term length).
- *Longest Common Subsequence Ratio*: A cognate feature measuring the longest common non-consecutive sequence of characters between two strings
- *isFirstWordCovered*: A combined feature indicating whether the first word in the source

[1] For combined features, a word is considered as covered if it can be found in the corresponding set of Giza++ translations or if one of the cognate-based measures (Longest Common Subsequence, Longest Common Substring, Levensthein Distance, Needleman-Wunsch Distance, Dice) is 0.70 or higher (set experimentally by (Aker et al., 2013b))

term has a translation or transliteration in the target term.

- *isFirstWordCognate*: a binary feature which returns True if the longest common consecutive string (LCST) of the first words in the source and target terms divided by the length of the longest of the two words is greater than or equal to a threshold value of 0.7 and both words are longer than 3 characters.

Repar et al. (2019) start by reproducing this approach, but were unable to replicate the results. During the subsequent investigation, they discovered that using the same balance ratio in the training and test sets (i.e. 1:200, which was set by Aker et al. (2013b) to mimic real-world scenarios) have a significant impact on the performance of the algorithm. Furthermore, they filter training set term pairs based on term length and feature values (hence the different training set sizes in Table 2) and develop new cognate-based features.

The system requires several language-specific resources:

- A large parallel corpus to calculate word alignment probability with Giza++. The system in Repar et al. (2019) uses the DGT translation memory (Steinberger et al., 2013).
- A list of aligned terms that serve as training data. The system in Repar et al. (2019) uses the Eurovoc thesaurus (Steinberger et al., 2002). 600 Eurovoc term pairs are used as test data, while the rest is used for training.
- Transliteration rules for the construction of reverse cognate-based features (cognate features are constructed twice: first the target word is transliterated into the source language script, then the source word is transliterated in the target language script).

The constructed features are then used to train the SVM classifier which can be used to predict the alignment of terms between two languages.

4 Resources for the Estonian-Russian experiment

While the DGT translation memory and the Eurovoc thesaurus support all official EU languages, there is no Russian support since Russia is not an EU member state. In order to train the classifier, we therefore had to find alternative resources.

For the parallel corpus, we made experiments with the Estonian Open Parallel corpus[2] and the Estonian-Russian OpenSubtitles corpus from the Opus portal[3]. The OpenSubtitles corpus performed better, most likely due to its much larger size (85,449 parallel Estonian-Russian segments in the Estonian Open Parallel corpus vs. 7.1 million segments in the OpenSubtitles corpus).

While finding parallel Estonian-Russian corpora was trivial due the the list of available corpora on the Opus portal, finding an appropriate bilingual terminological database proved to be more difficult. Ideally, we would want to use a media or news-related Estonian-Russian terminological resource, but to the best of our knowledge, there was none available. Note that the terminological resource needs to have at least several thousand entries: the Eurovoc version used by Repar et al. (2019) contained 7,083 English-Slovene term pairs. We finally settled on the environmental thesaurus Gemet[4], which at the time had 3,721 Estonian-Russian term pairs. For the transliteration rules, we used the Python pip package transliterate [5] to generate the reverse dictionary-based features.

5 Results

Repar et al. (2019) ran a total of 10 parameter configurations. We selected three of those to test on the Estonian-Russian dataset. The first one is the configuration with a positive/negative ratio of 1:200 in the training set, which significantly improved recall compared to the reproduction of Aker et al. (2013b), the second one is the same configuration with additional term filtering, which was overall the best performing configuration in Repar et al. (2019), and the third one is the Cognates approach which should give greater weight to cognate words. As shown in Table 2, the overall results are considerably lower than the results in Repar et al. (2019), in particularly in terms of recall. One reason for this could be that the term filtering heuristics developed in Repar et al. (2019) may not work well for Estonian and Russian as they do for other languages. For example, 1.3 million candidate term pairs were constructed for the English-Slovene lan-

[2] https://doi.org/10.15155/9-00-0000-0000-0000-0002AL

[3] opus.nlpl.eu

[4] https://www.eionet.europa.eu/gemet/en/themes/

[5] https://pypi.org/project/transliterate/

No.	Config ET-RU	Training set size	Pos/Neg ratio	Precision	Recall	F-score
1	Training set 1:200	627,120	1:200	0.3237	0.2050	0.2510
2	Training set filtering 3	30,954	1:200	0.9000	0.0900	0.1636
3	Cognates approach	33,768	1:200	0.7313	0.0817	0.1469

Table 2: Results on the Estonian-Russian language pair. No. 1 presents the results of the configuration with a positive/negative ratio of 1:200 in the training set, no. 2 presents the results of the same configuration with additional term filtering, which was overall the best performing configuration in Repar et al. (2019), and No. 3 presents the results of the Cognates approach which should give greater weight to cognate words.

ET	RU	Evaluation
kontsert	концерт	exact match
kosmos	космос	exact match
majandus	экономика	exact match
juhiluba	водительские права	exact match
lõbustuspark	парк развлечений	exact match
unelmate pulm	свадьба	partial match
eesti mees	мужчина	partial match
indiaani horoskoop	гороскоп	partial match
hiina kapsas	капуста	partial match
hulkuvad koerad	собаки	partial match
eesti autospordi liit	эстонский футбольный союз	no match
Kalevi Kull	орел	no match
honda jazz	джаз	no match
tõnis mägi	гора	no match
linkin park	парк	no match

Table 3: Examples of exact, partial and no match tag pairs produced by the system.

guage pair and around one half of those were filtered out during the term filtering phase. On the other hand, only around 33,000 Estonian-Russian candidate pairs out of the total 627,000 survived the term fitering phase in these experiments. Another reason for the lower performance is likely the content of the language resources used to construct the features. Whereas Repar et al. (2019) use resources with similar content (EU legislation), here we have dictionary-based features constructed from a subtitle corpus and term pairs from an environmental thesaurus.

We then used the best performing configuration to try to align the Estonian and Russian tags from the dataset provided by Ekspress Meedia. The size of the dataset (59,632 Estonian tags and 6,198 Russian tags) and the fact that the system must test each possible pairing of source and target tags meant that the system generated around 370 million tag pair candidates which it then tried to classify as positive or negative. This task took more than two weeks to complete, but at the end it resulted in 4,989 positively classified Estonian-Russian tag pairs. A subset of these (500) were manually evaluated by a person with knowledge of both languages provided by Ekspress Meedia according to the following methodology:

- C: if the tag pair is a complete match
- P: if the tag pair is a partial match, i.e. when a multiword tag in one langauge is paired with a single word tag in the other language (e.g. eesti kontsert — концерт, or *Estonian concert — concert*)
- N: if the tag pair is a no match

Of the 500 positively classified tag pairs that were manually evaluated, 49% percent were deemed to be complete matches, a further 25% were evaluated as partial matches, and 26% were considered to be wrongly classified as positive tag pairs. The evaluator observed that "the most difficult thing was to separate people's names from toponyms, such as a famous local singer called "Tõnis Mägi", a district in Talinn called "Tõnismägi"

and a mountain named "Muna Mägi". More examples of exact, partial and no match alignments can be found in Table 3.

6 Conclusions and future work

In this paper, we reused an existing approach to terminology alignment by Repar et al. (2019) to align a set of Estonian and Russian tags provided by the media company Ekspress Meedia. The approach requires several bilingual resources to work and it was difficult to obtain relevant resources for the Estonian-Russian language pair. Given the domain of the tagset, i.e. news and media, the selected resources (subtitle translations and an environmental thesaurus) were less than ideal. Nevertheless, the approach provided respectable results with 74% of the positive tag pairs evaluated to be at least a partial match.

When asssessing the performance of the approach, one has to take into account the fact that the tagset is heavily unbalanced with almost 60,000 Estonian tags compared to a little over 6,000 Russian tags. This means that for many Estonian tags, a true equivalent was simply not available in the tagset.

For future work, we plan to integrate additional features into the algorithm, such as those based on novel neural network embeddings which may uncover additional hidden correlations between expressions in two different languages and may provide an alternative to large parallel corpora which are currently needed for the system for work. In terms of the Estonian and Russian language pair, additional improvements could be provided by taking into account the compound-like structure of many Estonian words. Finally, we will look into techniques that would allow us to pre-filter the initial list of tag pairs to reduce the total processing time.

References

Ahmet Aker, Monica Paramita, and Rob Gaizauskas. 2013a. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–411, Sofia, Bulgaria. Association for Computational Linguistics.

Ahmet Aker, Monica Paramita, and Rob Gaizauskas. 2013b. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 402–411.

Robert Gaizauskas, Ahmet Aker, and Robert Yang Feng. 2012. Automatic bilingual phrase extraction from comparable corpora. In *24th International Conference on Computational Linguistics*, pages 23–32.

Thorsten Joachims. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Andraž Repar, Matej Martinc, and Senja Pollak. 2019. Reproduction, replication, analysis and adaptation of a term alignment approach. *Language Resources and Evaluation*, pages 1–34.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2013. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*.

Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, pages 101–121.

Chapter 4

Monolingual Terminology Extraction

In this chapter, we describe our research in the area of monolingual terminology extraction. It is divided into three sections. We first describe the dataset used for both approaches in Section 4.1 and use the next two sections to describe a machine-learning approach and a sequence-labeling approach to terminology extraction, respectively. The paper relevant for the first approach was presented at the 18th TOTh International Conference in 2024 and will be published in its proceedings. An extended version of the paper is also available on Arxiv.org:

Repar, A., Lavrač, N., & Pollak, S. (2022). Extracting domain-specific terms using contextual word embeddings. arXiv preprint. Available: https://arxiv.org/abs/2502.17278
Code available here: https://github.com/andrazrepar/term-embeddings.

The relevant papers for the second approach are:

Tran, H., Martinc, M., Repar, A., Doucet, A., & Pollak, S. (2022). A Transformer-based sequence-labeling approach to the Slovenian cross-domain automatic term extraction. Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia, 196-204.
Code available here: https://github.com/honghanhh/sdjt-ate.

Tran, H., Martinc, M., Repar, A., Ljubešić, N., Doucet, A., & Pollak, S. (2024). Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling?. Machine Learning, 113, 1-30.
Code available here: https://github.com/honghanhh/ate_nobi.git.

4.1 Dataset

The RSDO5 corpus (Jemec Tomažin et al., 2021) was compiled for developing and evaluating terminology extraction methods in Slovene and was inspired by the ACTER corpus (Rigouts Terryn, Hoste, & Lefever, 2020b). It consists of 12 texts with 250,000 words and 38,000 manually annotated terms[1]. The corpus texts, published between 2000 and 2019, belong to the fields of biomechanics, linguistics, chemistry, or veterinary science. For each domain, they include: a PhD thesis, a graduate level text book, and a journal article. The entire texts were annotated for terminology and allow for evaluation of methods in terms of precision and recall. Apart from the manually annotated terms, the corpus was automatically annotated, i.e. performing tokenization, sentence segmentation, lemmatization, assigning morphological features and dependency syntax using the Classla pipeline

[1]Note that this number refers to all occurrences of all terms.

Table 4.1: Number of lemmatized terms, terms with frequency of 1, terms with frequency of 2 and terms with frequency of more than 2, per domain in the RSDO5 corpus.

domain	lemma terms	freq=1	freq=2	freq>2
biomechanics	1,596	891	266	439
linguistics	3,102	2,115	415	532
veterinary	1,580	880	245	455
chemistry	3,379	2,098	483	798

(Ljubešić & Dobrovoljc, 2019). It is available in the Clarin.si repository.[2]

4.2 A Machine-learning Approach to Monolingual Terminology Extraction

In this section, we describe a terminology extraction methodology that combines two traditional aspects of automated terminology extraction with a novel contextual-embedding approach in a machine learning setting. Focusing on the Slovenian language, which is an under-resourced Slavic language with a rich morphology, we conducted the first experiments on a new RSDO5 (see Section 4.1) corpus of term-annotated texts created as part of the RSDO national project. The proposed approach starts with the corpus — we first analyze the annotated terms in the corpus and study their part-of-speech tags. While traditional systems use a set of pre-defined part-of-speech patterns to identify the initial candidate terms (CTs), we take a different approach and instead define a shallow filter for CTs, which considers only very basic part-of-speech based information. In our approach, we generate three types of features (linguistic, statistical and contextual) and use them to train a linear support vector machine (SVM) classifier. Since the RSDO5 corpus contains four different domains, we train the algorithm on three domains and test its performance on the fourth domain using the standard measures of precision, recall and the F_1 score (as in the related work by Rigouts Terryn, Hoste, Drouin, et al. (2020)).

4.2.1 Corpus analysis

Table 4.1 contains the basis statistics of the terms in the RSDO5 corpus described in Section 4.1 relevant for our term extraction approach. It contains a total of 9,657 unique lemma term forms and a large number of these occur only once or twice in an individual domain of the corpus. We have analyzed the annotated terms in the corpus with respect to token and character lengths, POS tags of unigram terms, POS tags of the first token in the terms, POS tags of the last token in the terms, and frequency of different POS tags in the terms.

As can be observed in Figure 4.1, most annotated terms have 4 or less tokens. The longest term per domain has 6 tokens in the chemistry domain, 11 tokens in the biomechanics domain, 10 tokens in the veterinary domain and 8 tokens in the linguistics domain. The vast majority of terms also have more than 4 characters (see Figure 4.2). We can also observe that[3]:

[2]https://www.clarin.si/repository/xmlui/handle/11356/1400

[3]While Figures 4.1 and 4.2 are produced based on the lemmatized term lists, in the counts of Figures 4.3, 4.4, 4.5 all different appearances of terms are considered 4.6, due the fact that the syntactic parsing algorithm (Classla) could produce different annotations in different contexts.

- almost all unigram terms are either nouns (NOUN) or proper nouns (PROPN) (for details, see Figure 4.3),
- most terms start with either an adjective (ADJ), noun (NOUN) or proper noun (PROPN) (for details, see Figure 4.4),
- most multi-word unit terms end with a noun (NOUN) or a proper noun (PROPN) (for details, see Figure 4.5),
- nouns (NOUN) and adjectives (ADJ) are by far the most frequent POS tags that appear in the terms, but adverbs (ADV), adpositions (ADP) and proper nouns (PROPN) can also be found; other POS tags occasionally appear in some terms, but the number of occurrences is low and may in some cases be attributed to errors during the syntactic parsing process (for details, see Figure 4.6).

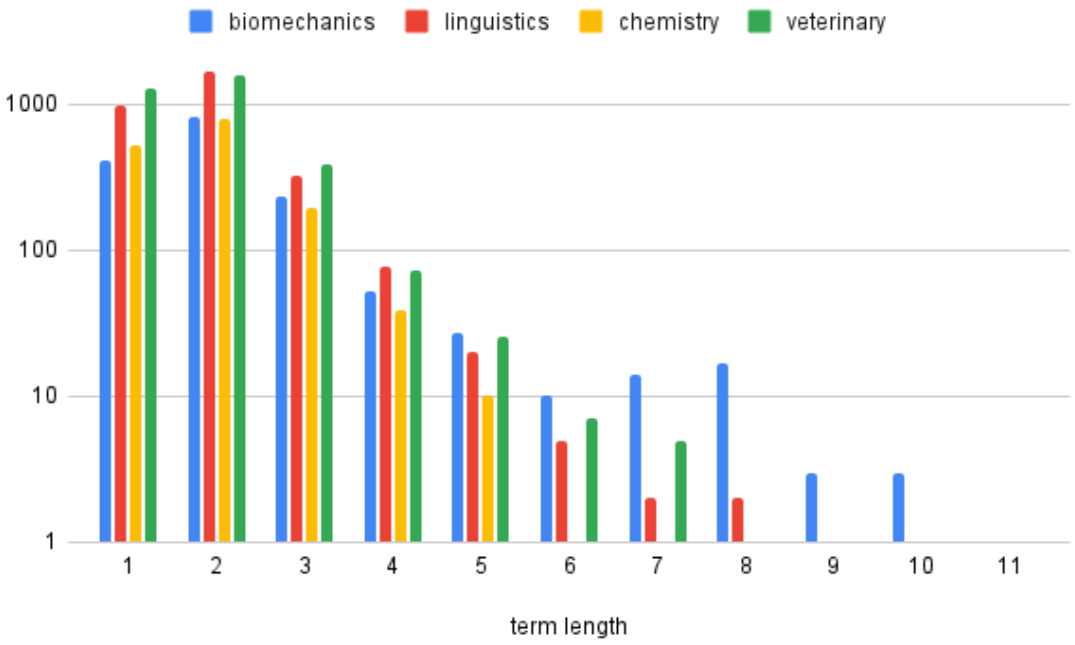

Figure 4.1: Number of tokens in gold standard (lemmatized) terms.

4.2.2 System overview

This section describes the architecture of the system. We use a machine learning approach to automated terminology extraction (ATE), which means that for each term candidate, we generate a set of features and then use the corresponding labels for model training. Our approach is similar to the ones developed by Ljubešić et al. (2019) and Rigouts Terryn et al. (2021a) in some respects, but it differs from them in two major aspects: 1) instead of using a pre-defined set of linguistic patterns to identify CTs, we apply 6 filters to all possible n-grams in the input corpus up to a user-defined length n, and 2)in addition to linguistic and statistical features, we also employ a set of novel contextual features based on ELMo (Peters et al., 2018) contextual word-embeddings. A general overview of the system is depicted in Figure 4.7.

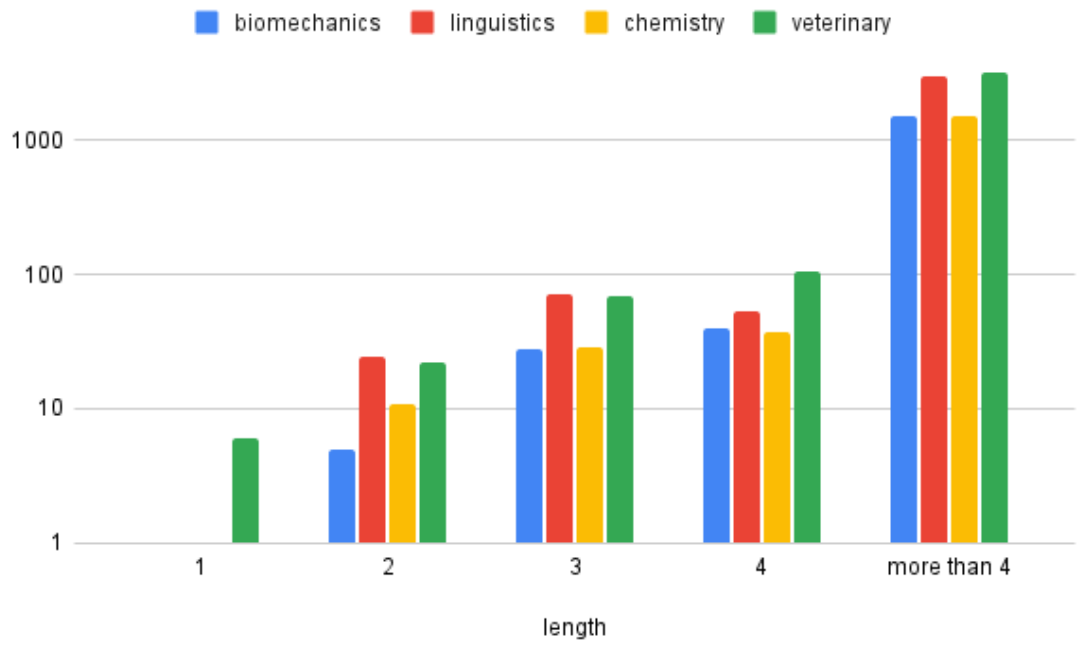

Figure 4.2: Character length of gold standard (lemmatized) terms (including spaces for MWUs).

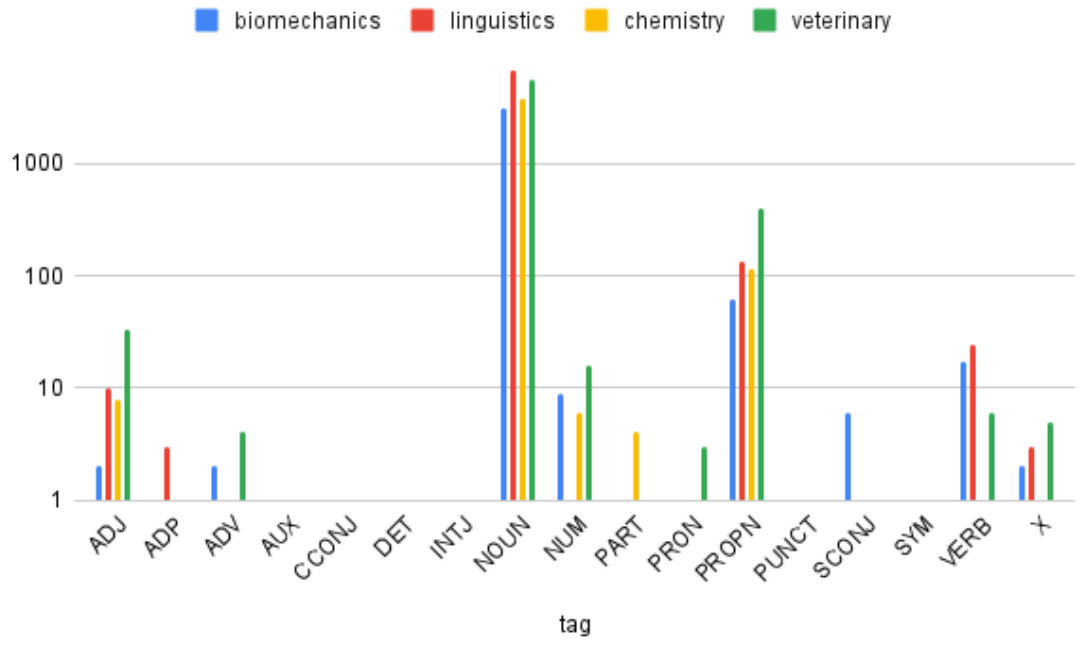

Figure 4.3: POS tag of gold standard unigram (non-lemmatized) terms.

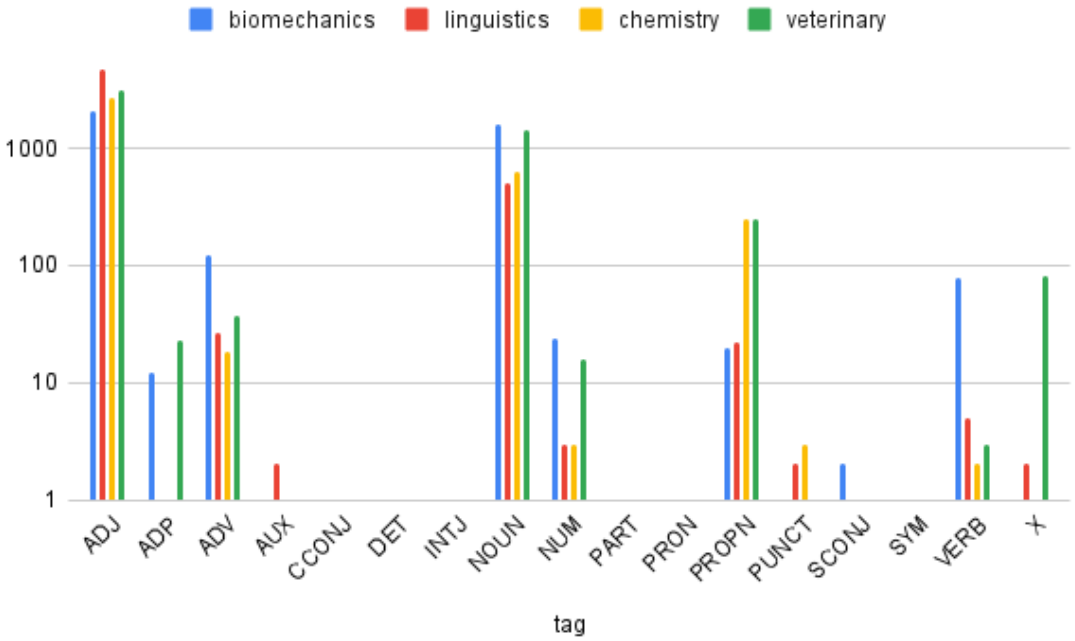

Figure 4.4: First POS tag of annotated (non-lemmatized) terms.

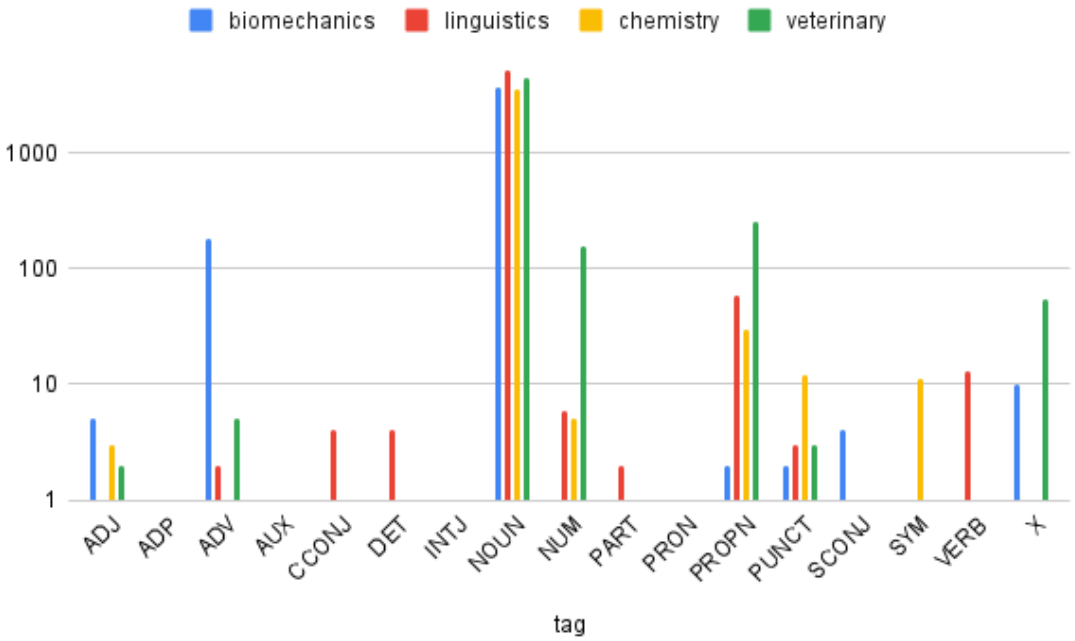

Figure 4.5: Last POS tag of annotated (non-lemmatized) terms (longer than 1 token).

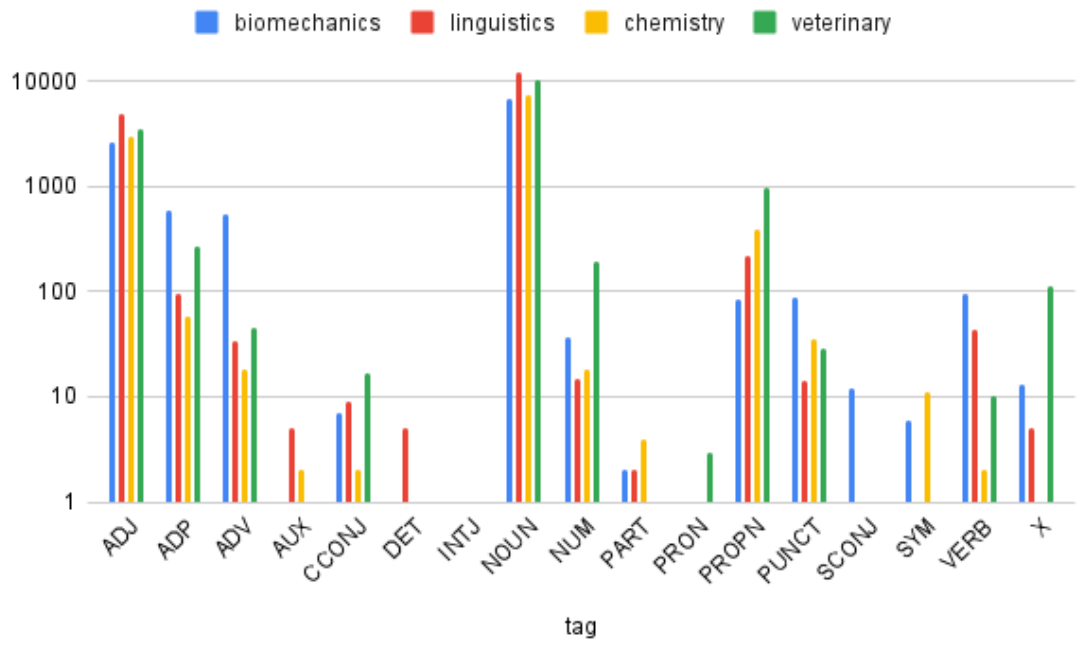

Figure 4.6: Frequency of POS tags that appear in the annotated (non-lemmatized) terms.

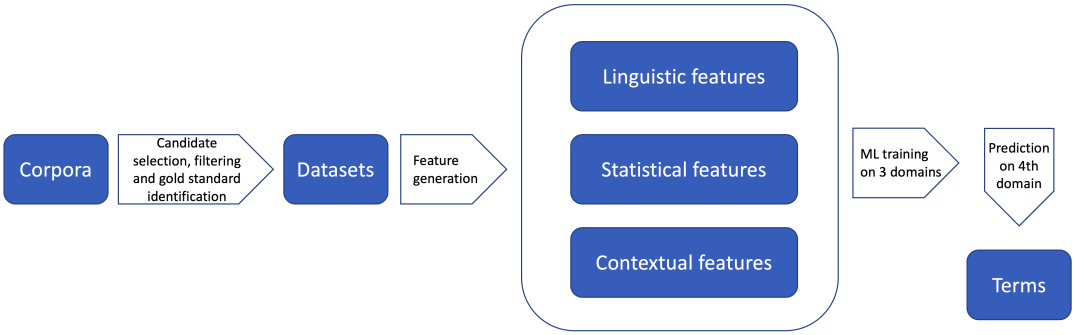

Figure 4.7: Overview of the terminology extraction machine learning system using different types of features. Starting with 4 domain corpora, we generate the datasets by identifying gold standard terms and generating candidates and then calculating 3 types of features. 3 datasets are used to train a machine learning model, while the 4th is used to predict the terms for evaluation.

4.2.3 Dataset pre-processing

In traditional ATE systems, candidate terms (CTs) are usually selected based on a pre-defined list of POS patterns. For example, all adjective-noun (ADJ+NOUN) sequences, such as "supervised learning" or "basic rule", are considered CTs. Since most terms, in particular high frequent ones, correspond to one of the standard patterns, this allows us to quickly filter out a large number of low-quality CTs. However, defining a POS pattern list that would effectively cover terms across various types of domains is difficult. While some common patterns (e.g., NOUN+NOUN or ADJ+NOUN) can be considered universal, other patterns may be domain-specific and maintaining different pattern lists for many different domains can be cumbersome. Using a traditional pattern-based approach, we also discard potentially valid terms that do not correspond to one of the POS patterns either because they follow a non-standard POS pattern or because they are unusually long. Since most patterns are rarely longer than 4 or 5 tokens (see Figure 4.1), longer terms would automatically be discarded.

Instead of relying on a list of POS patterns to identify CTs we apply a shallow filter to all n-grams up to a pre-defined maximum length value. In our case, we set the maximum length to 11, since this is the longest annotated term in the RSDO5 corpus (see Figure 4.1). The shallow filter is based on the analysis of the terms in the RSDO5 corpus and describes the general linguistic characteristics of the terms. The rules are described in detail below:

1. *Terms have to be longer than 3 characters.*
   As evident from Figure 4.2, the vast majority of terms (97.87% in the biomechanics domain, 96.91% in the linguistics domain, 97.47% in the chemistry domain and 97.10% in the veterinary domain) are longer than 3 characters.

2. *Only nouns can be single word terms.*
   As evident from Figure 4.3, the vast majority of unigram terms (98.78% in the biomechanics domain, 99.40% in the linguistics domain, 99.50% in the chemistry domain and 98.82% in the veterinary domain) are either nouns (NOUN) or proper nouns (PROPN)[4].

3. *Patterns longer than 1 have to end with a noun (NOUN) or proper noun (PROPN) to be terms.*
   As evident from Figure 4.5, the vast majority of last POS tags of annotated terms, longer than one token (94.92% in the biomechanics domain, 99.34% in the linguistics domain, 99.22% in the chemistry domain and 95.54% in the veterinary domain) are either nouns (NOUN) or proper nouns (PROPN).

4. *Patterns not starting with adjectives (ADJ), adverbs (ADV) or nouns (NOUN, PROPN) are not terms.*
   As evident from Figure 4.4, the vast majority of first POS tags of annotated terms (96.98% in the biomechanics domain, 99.74% in the linguistics domain, 99.72% in the chemistry domain and 97.51% in the veterinary domain) are either adjectives (ADJ), adverbs (ADV), nouns (NOUN) or proper nouns (PROPN).

5. *If a pattern contains a verb (VERB), a symbol (SYM), a subordinating conjunction (SCONJ), punctuation (PUNCT), a pronoun (PRON), a particle (PART), an interjection (INTJ), a determiner (DET), a coordinating conjuction (CCONJ), an*

[4]Note that we do not distinguish between nouns and proper nouns, because we found that the syntactic parsing process is unreliable when it comes to nouns that can be both regular nouns and proper nouns (such as the word "commission" which can be used in the general sense or as part of the proper name "European Commission").

Table 4.2: Dataset filtering effects, where GS denotes gold standard.

	GS terms	Filtered out	Max. recall	Candidates
biomechanics	1,596	138	0.91	12,847
linguistics	3,102	277	0.91	22,610
chemistry	1,580	115	0.93	15,417
veterinary	3,379	481	0.86	17,996

auxiliary verb (AUX) or other (X), it is not a term.
As evident from Figure 4.4, only a small fraction of annotated terms (2.01% in the biomechanics domain, 0.48% in the linguistics domain, 0.51% in the chemistry domain and 1.12% in the veterinary domain) contain any of these POS tags. Despite the fact that adpositions (ADP) and adverbs (ADV) feature in a quite significant number of CTs, in particular in the biomechanics domain (10.16%)[5], we discovered during error analysis that a large number of wrongly predicted terms contain adverbs and/or adpositions.

6. *If term contains a comma or an underscore, it is not a term.*

Using these 6 filters, we are able to significantly reduce the number of CTs while maintaining adequate gold standard coverage. The maximum recall per domain is 0.91 for the biomechanics and linguistics domain, 0.86 for the veterinary domain and 0.93 for the chemistry domain, while the number of CTs (including valid gold standard terms) is 12,847 for the biomechanics domain, 22,610 for the linguistics domain, 17,996 for the veterinary domain and 15,417 for the chemistry domain. For details, see Table 4.2 showing the total number of gold standard terms being filtered out.

4.2.4 Feature construction

Similar to traditional ATE systems, we generate linguistic and statistical features and then we also add a third category of features that are based on contextual word embeddings.

4.2.4.1 Linguistic features

Contrary to some traditional systems, such as the one developed by Justeson and Katz (1995), we do not use pre-defined linguistic patterns, but instead generate features based on a set of loosely defined pattern rules utilizing the Universal Dependency (UD) POS tags (de Marneffe et al., 2021), to avoid overlooking terms due to missing patterns as explained in Section 4.2.3.

We generate several vectors of UD tags for each CT depending on the UD part-of-speech value as described below. Each vector has a length of 17 corresponding to the number of all possible UD tags:

- *StartUD*: a one-hot vector where the UD tag of the first token in the CT has a value of 1, while the rest have a value of 0,
- *EndUD*: a one-hot vector where the UD tag of the last token in the CT has a value of 1, while the rest have a value of 0; in the case of unigram CTs, the *StartUD* and *EndUD* vectors are the same,

[5]They are less prevalent in other domains: 0.74% in linguistics, 0.69% in chemistry and 2.05% in veterinary.

Table 4.3: The four vectors generated during feature construction for the term *supervised machine learning* annotated with the following UD tags: ADJ NOUN NOUN.

Vector	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
StartUD	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EndUD	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
AnywhereUD	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CountOfUD	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0

- *AnywhereUD*: a vector where the UD tags that appear anywhere in the CT other than the first and last position have a value of 1, while the rest have a value of 0; in the case of unigram and bigram CTs, this vector has only zero values,
- *CountOfUD*: a vector indicating the number of occurrences of each UD tag anywhere in the CT.

For an illustration of the vectors generated, see Table 4.3. Finally, we also generate an additional numeric feature that counts the number of unique POS tags in the term candidate:

- *NoUniquePos*

The vectors are concatenated, resulting in a representation of 69 features: 51 features with binary values, and 18 (17 from CountOfUD and 1 NoUniquePos) with numeric values.

4.2.4.2 Statistical features

For statistical features, we use the *termhood* measure from Vintar (2010), which is based on the premise that domain-specific terms are used more frequently (in relative terms) in domain texts than in the general language. But instead of calculating a single termhood value, we generate the three core variables (general corpus frequency, domain corpus frequency and term length) from the termhood formula as separate features[6]:

- *TermGenFreq*: the sum of the relative frequencies in a general corpus of individual tokens constituing a CT,
- *TermDomFreq*: the sum of the relative frequencies in a domain-specific corpus of individual tokens constituing a CT,
- *TermLength*: which corresponds to the length of the CT (i.e. number of tokens).

For calculating general corpus relative frequency, we used a word frequency list from the Gigafida 2.0 Slovene reference corpus (Krek et al., 2020), whereas the domain corpus is the training data from the RSDO5 corpus described in Section 4.1. The sums of relative frequencies are calculated using the following formula:

$$\sum_{1}^{n} \log \frac{f_n}{N} \tag{4.1}$$

where n represents the number of tokens in the CT, f_{n} is the frequency of each token in the CT and N is the size of the corpus in tokens. The same formula is used for the calculation of general and domain-specific corpora relative frequencies. The features are concatenated to the linguistic feature vector.

[6]Note that we use the following stoplist to exclude words from frequency calculations: *brez, do, iz, z, s, za, h, k, proti, kljub, čez, skozi, zoper, po, o, pri, po, z, s, na, ob, v, med, nad, pod, pred, za*

4.2.4.3 Contextual features

To generate contextual features, which is also the main novelty of our approach, we utilize a premise that is similar to statistical termhood measures. Just as termhood suggests that domain-specific terms are used more frequently in domain-specific corpora than in general corpora, so we hypothesize that domain-specific terms are used in different contexts in domain-specific corpora compared to general corpora.

For the calculation of contextual embeddings, we used the ELMo model for Slovenian created by Ulčar (2019), which was trained on the Gigafida 2.0 corpus for 10 epochs. General corpus contextual embeddings were calculated for the top 200k most frequent tokens from the publicly available ccGigafida corpus (Logar et al., 2013). To produce the values, the first LSTM layer was used (based on Reimers and Gurevych (2019a) who report that the middle layer is the best single layer and comparable to concatenating all three layers) to produce the vector values. Each instance of a word has its own vector, based on the context it appears in. These vectors have been averaged, so that each word has only one corresponding vector representing its average context in the general corpus. We then calculate average word embeddings for every word in the domain corpus. To do this, we first tokenize the corpus into sentences and then generate word embeddings for every sentence using the AllenNLP Python library (Gardner et al., 2018) using the same ELMo settings as for the generation of the general domain corpus embeddings. We iterate over every word in the sentence and calculate the average embedding of its lemma in the domain corpus by adding up all vectors and dividing them by the number of occurrences of the lemma in the corpus. While for single word terms the average lemma embedding is the final representation, for every multi-word term, we calculate the average embedding by summing up the average lemma embeddings of all tokens in the term and dividing the sum with the number of tokens in the term. All 1,024 dimensions of the resulting vector are then added to the dataset as features. In a similar manner, we then also generate the average general domain term embedding by summing up average lemma embeddings in the general corpus and dividing the sum with the number of tokens in the term. All 1,024 dimensions of the resulting vector (*elmo* feature vector) are again added to the representation feature vector[7].

The described approach of averaging contextual embeddings may lead to the loss of certain contextual information, causing the resulting averages to resemble static embeddings. To address this limitation, we implemented three additional features leveraging contextual embeddings:

- *elmoSim*, which is the cosine similarity of the domain-specific and general term embeddings calculated as described above, the motivation being that since true terms are used in different contexts in domain-specific and general language texts, the similarity between the two vectors would be smaller for true terms compared to expressions that are not valid terms.

- *elmoTermSim*, which is the cosine similarity of the domain-specific term embedding and the embedding of a seed term defined by the user[8], the motivation being that true terms would be used in similar contexts to a term representative of the domain.

[7]Future work could explore dimensionality reduction methods like Uniform Manifold Approximation and Projection (UMAP) to determine whether a lower-dimensional representation retains the essential contextual information while reducing redundancy.

[8]In the case of the RSDO dataset, we used the domain names as seed terms — "veterina" for the veterinary domain, "jezikoslovje" for the linguistics domain, "biomehanika" for the biomechanics domain and "kemija" for the "chemistry" domain.

Table 4.4: F_1 scores of various algoritms across domain. Support vector machine has the highest average F_1 score.

Algorithm	bim	ling	chem	vet	average
decision tree	0.388	0.397	0.390	0.422	0.399
random forest	0.294	0.298	0.362	0.388	0.336
multiple layer perceptron	**0.536**	0.522	0.548	0.561	0.542
logistic regression	0.535	0.568	**0.563**	0.579	0.561
support vector machine	0.530	**0.569**	0.561	**0.594**	**0.564**

- *elmoStDev*, which is calculated as follows: for every lemma in the domain-specific corpus, we calculate the standard deviation of all its contextual embeddings and then for each term, we sum up the standard deviations of all lemmas and divide the sum with the number of tokens in the term, the motivation being that true terms in most cases appear in similar contexts within domain-specific texts which would result in smaller standard deviation compared to non-valid terms.

4.2.5 Experiments and results

4.2.5.1 Experimental setup

For model training, we experimented with five algoritms from the *sklearn* Python library. The best F_1 score was achieved by a SVM binary classifier with a linear kernel using the default settings (c=1). For detailed results, see Table 4.4. Following the setup proposed by Hazem et al. (2020), we use three domains for training and one for testing, which means that we run the experiments four times with a different domain used for testing each time. Evaluation is performed using the standard measures of precision, recall and F_1 scores. Precision is calculated as the number of true positive terms divided by the number of all predicted terms, recall is calculated as the number of true positive terms divided by the number of GS terms and F_1 score is calculated as the harmonic mean of precision and recall.

We compare the results to the state-of-the-art for Slovenian (Ljubešić et al., 2019) (the code for that approach is freely available[9]), to the LUIZ approach by Vintar (2010) as implemented in Repar et al. (2019), where a joint list of single and multi-word terms is produced, and to an approach where we use corpus-based patterns similar to Rigouts Terryn et al. (2021a) instead of our filtering rules[10]. For Ljubešić et al. (2019), we trained the MWU and SWU models for all four combinations of training and testing domains and evaluated their performance against the gold standard terms. Minimum frequency was set to 1, which is the same as in our approach. For the LUIZ approach, which does not classify the terms but produces a ranked list of term candidates, we used a cuttoff which corresponded to the number of terms predicted by our approach for a specific domain (i.e. if our approach predicted n terms for a domain, we considered the *top* n terms according to the LUIZ score for this domain).

4.2.5.2 Results

With our approach, we achieve F_1 scores of 0.530 for the biomechanics domain, 0.569 for the linguistics domain, 0.561 for the chemistry domain and 0.594 for the veterinary domain (see

[9]https://github.com/clarinsi/kas-term

[10]We first collected all patterns of the terms annotated in the RSDO corpus and then generated candidates that correspond to these patterns.

Table 4.5: Precision, recall and F_1 score compared to competitive approaches for Slovenian by Vintar (2010) and Ljubešić et al. (2019). *Our approach* uses the shallow filter described in 4.2.3, whereas *Pattern approach* uses corpus-based patterns similar to Rigouts Terryn et al. (2021a).

Model	Test	Precision	Recall	F_1 score
Our approach	bim	0.650	**0.448**	**0.530**
Pattern approach	bim	**0.694**	0.342	0.458
LUIZ	bim	0.359	0.393	0.363
Ljubešić et al. (2019)	bim	0.538	0.248	0.339
Our approach	ling	0.672	**0.494**	**0.569**
Pattern approach	ling	**0.678**	0.446	0.538
LUIZ	ling	0.338	0.393	0.363
Ljubešić et al. (2019)	ling	0.522	0.254	0.341
Our approach	chem	0.691	**0.472**	**0.561**
Pattern approach	chem	**0.694**	0.374	0.486
LUIZ	chem	0.239	0.444	0.311
Ljubešić et al. (2019)	chem	0.478	0.314	0.378
Our approach	vet	**0.688**	**0.523**	**0.594**
Pattern approach	vet	0.670	0.487	0.564
LUIZ	vet	0.400	0.349	0.373
Ljubešić et al. (2019)	vet	0.669	0.193	0.299

Table 4.5). These results are comparable with several strong baselines, including results by Rigouts Terryn et al. (2021a) and Lang et al. (2021), but lower than the sequence labelling approach presented in Section 4.3. Moreover, there is very little variation between domains.

We also see that our approach exceeds the approach by Ljubešić et al. (2019) in both precision and recall, which is not surprising given that their method relies heavily on frequency-based features, which work best with high frequency CTs, as well as the more traditional LUIZ approach.

For a discussion on the practical usability of these results, see Section 5.2.

4.2.5.3 Ablation study

We wanted to analyze the impact of different feature types described in Section 4.2.4 on the final results. One approach, particularly often used in the evaluation of deep learning algorithms, is ablation study (Meyes et al., 2019). Analogous to ablation in biology, ablation in machine learning denotes the removal of individual components and studying the effect on the results. In line with this, we have removed the individual feature types from the dataset one-by-one and analyzed the results available in Table 4.6.

We can observe that removing each feature type does reduce the F_1 scores in all domains and removing both statistical and linguistic features results in an even bigger drop in F_1 scores. When we removed the statistical features but kept linguistic and contextual features, we observed a drop in F_1 score performance by 11.70% in the biomechanics domain, 7.91% in the linguistics domain, 13.37% in the chemistry domain and 5.72% in the veterinary domain. When we removed the pattern features but kept statistical and contextual features, we observed a drop in F_1 score performance by 18.30% in the biomechanics domain, 13.18% in the linguistics domain, 11.59% in the chemistry domain and 9.43% in the veterinary domain. When we removed both statistical and linguistic features but kept contextual features, we observed a drop in F_1 score performance by 34.34% in the

Table 4.6: Ablation study results showing F_1 scores with different combinations of feature types (C — Contextual, P — Pattern, S — Statistical).

Test domain	C&P&S	C&P	C&S	S&P	C	S	P
bim	**0.530**	0.468	0.433	0.206	0.348	0.000	0.003
ling	**0.569**	0.524	0.494	0.174	0.412	0.000	0.000
chem	**0.561**	0.486	0.496	0.247	0.418	0.000	0.001
vet	**0.594**	0.560	0.538	0.089	0.489	0.000	0.000

biomechanics domain, 27.59% in the linguistics domain, 25.49% in the chemistry domain and 17.68% in the veterinary domain. Using only statistical and pattern features, either together or independently, produces almost no correct predictions.

All three different sets of features contribute to the final result. The drop in F_1 score performance when removing linguistic features is somewhat larger than when removing statistical features (with the exception of the chemistry domain) and when we remove both statistical and linguistic features, the results are even worse. However, the results when using just contextual features are still respectable, in particular in terms of precision, which is above 0.630 for all four domains.

4.2.5.4 Error analysis

We performed error analysis of the results obtained with the best performing model (i.e. the model based on all three feature types described in Section 4.2.4). When looking at the false positive predictions in all four domains, we were immediately reminded of the issue we mentioned in the introduction, namely that there is no clear definition of the nature of domain-specific terms. On first look and with the caveat that we are not experts in any of these domains (with the possible exception of linguistics), it would seem that many of the false positives could be valid terms. For example[11], *atletska steza (running track)*, *živčni končič (nerve ending)* and *upogibalka (flexor)* in the biomechanics domain, *jezikoslovni model (linguistic model)*, *kodna tabela (code table)* and *nacionalni korpus (national corpus)* in the linguistic domain, *prekurzor (precursor)*, *spektroskopija, (spectroscopy)* and *chronbachov koeficient (cronbach coefficient)* in the chemistry domain and *žvekalka (masseter muscle)*, *stomatitis (stomatitis)* and *nekrotično vnetje (necrotic inflammation)* in the veterinary domain could to an untrained eye look like valid domain-specific terms. We also believe that some of the false positive predicted terms (and many others) would also be useful at least in a "semi-automatic" terminology extraction setting, where CTs are first extracted automatically and then evaluated by a domain expert.

In addition to the terms described above, we noted two additional issues among false positives. The first one is general terms/words being predicted as terms, such as leto (year), mesto (city), sistem (system), stopnja (rate), proces (process), sprememba (change), skupina (group), delo (work), zbirka (collection), primer (example), pogoj (condition) etc. The reason for these wrongly predicted terms could again be related to the unclear nature of domain-specific terms, because the gold standard contains some terms that, on first look, do not look much different than the ones mentioned above, such as sila (force), enota (unit), sejem (fair), komora (chamber) etc. The second identified issue is related to the lemmatization algorithm in the Classla pipeline. For example, false positives contains wrongly lemmatized CTs, such as "regrgrposs" (lemma of "REGR_r_OsSU") and "doogpodlaht"

[11]Since Slovenian is not a widely known language, we provide English translations in brackets. In addition, please note that while we use canonical forms in the examples for better readability, the system actually produces lemmatized forms.

(lemma of "D_O_podlahti") or "mehanovsprejemnik" (lemma of "mehano-sprejemniki") and "skorajstrokovnjak" (lemma of "skoraj-strokovnjakov").

4.2.6 Conclusions

This section presents a machine learning terminology extraction system, which combines elements of traditional termhood- and unithood-based systems with a novel contextual-word-embedding-based approach that takes advantage of the differences in the domain-specific and general language contexts which terms appear in. In addition, we also introduce a novel method of candidate term selection — instead of being limited to a pre-defined list of part-of-speech patterns, we employ a shallow filter that offers greater flexibility in candidate term selection, in particular when it comes to unseen data and when implementing term extraction in user-facing applications[12]

We evaluated the system on a new corpus of term-annotated texts RSDO5 1.0 for Slovenian, created as part of the work in the Slovenian national language technology project RSDO. We compared the results to the existing state-of-the-art approach for Slovenian and were able to improve F_1 scores by a considerable margin in all four domains. The novel contextual features based on the ELMo embeddings appear to work well even for low-frequency terms (a well-known issue of traditional statistical methods, many of which are based on frequency counts). In addition, the results also exhibit little variation between test domains.

In terms of time, the most time-consuming part is the calculation of contextual embeddings on the large reference corpus. But since this can be computed only once and reused for further calculations, the system is fairly quick for a corpus of modest size (i.e. 50 to 100 thousand words). E.g., calculating domain-specific embeddings and applying the model is performed in approximately half an hour on a standard laptop without specialized machine learning hardware. This is an acceptable setting for practical applications, e.g. in translation industry or terminology dictionary construction setting. While the current version works only for Slovenian, it would be relatively easy to adapt it to other languages, provided that a suitable general language corpus is available.

4.3 A Sequence-labeling Approach to Monolingual Terminology Extraction

This section describes my involvement in the development of a sequence-labeling approach to monolingual terminology extraction where I focused mostly on the evaluation and analysis of the results and the design of the Slovenian annotated dataset. The relevant papers for this section is:

Tran, H., Martinc, M., Repar, A., Doucet, A., & Pollak, S. (2022). A Transformer-based sequence-labeling approach to the Slovenian cross-domain automatic term extraction. Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia, 196-204.
Tran, Hanh & Martinc, Matej & Repar, Andraz & Ljubešić, Nikola & Doucet, Antoine & Pollak, Senja. (2024). Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling?. Machine Learning, 113, 1-30.

[12]This research was conducted as part of a language technology project in Slovenia, where one the applications being in developed is a terminology portal with support for terminology extraction. We believe that non-linguist users would have difficulty defining a comprehensive pattern set and would rather work with verbal constructions such as "my term should start with POS1 and end with POS2", which correspond nicely with our filtering rules.

4.3.1 Description of the approach

Terminology extraction is treated as a sequence-labeling task where the model returns a label for each token in a text sequence using two different labeling regimes: the standard BIO scheme, as in Lang et al. (2021) and Rigouts Terryn et al. (2021a), where each token is annotated as being in the beginning of a term (B), inside a term (I) or outside a term (O), and NOBI, a novel labeling regime with two additional labels BN and IN, referring to a word being in the beginning or inside a nested term. For both labeling regimes, we experiment with XLMR (Conneau et al., 2020), a Transformer-based model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. The model is first trained to predict a label for each token in the input text sequence (e.g., we model the task as token classification) and then applied to the unseen text (test data). Finally, from the tokens or token sequences labeled as terms, the final candidate term list is composed.

The developed approach is evaluated in three settings:

- **Monolingual**, where there is a match between the language of the train set and the language of the test set
- **Crosslingual**, where the model is trained in one or more languages and tested on another language not appearing in the train set
- **Multilingual**, where the model is trained and tested on multiple languages

All three settings are applied in a cross-domain evaluation scenario, where we use two domains for training, another domain for validation, and the rest for testing.

4.3.2 Results

We evaluated the performance of the system by comparing the candidate list extracted from the test set with the manually annotated gold standard term list for that specific test set. We used exact string matching to compare the retrieved terms to the ones in the gold standard and calculate Precision (P), Recall (R), and F_1-score (F_1). In addition to the RSDO5 corpus described in Section 4.1, we also used ACTER (Rigouts Terryn, Hoste, & Lefever, 2020b), a corpus with 4 domains in three languages, which was created specifically for terminology extraction and on which the RSDO5 corpus was modeled.

On the ACTER dataset, results indicate that the cross-lingual and multilingual models in both versions of test data in most cases surpass the performance of the monolingual ones according to all evaluation metrics. Multilingual models also tend to outperform cross-lingual ones in F_1, but perform worse in terms of precision compared to monolingual and crosslingual ones. The proposed approaches significantly improve on benchmark approaches from the Termeval competition (Rigouts Terryn, Hoste, Drouin, et al., 2020) and others, such as Rigouts Terryn et al. (2021a) and Lang et al. (2021). On the RSDO5 dataset, the monolingual settings exhibit slightly better performance then multilingual ones, but the differences are small and could be attributed to the number of terms and length of terms per domain, as well as different combinations of training, validation and test sets. The performance on the RSDO5 corpus was also compared to the state-of-the-art approach for Slovenian by Ljubešić et al. (2019), which has been reimplemented using the data from the RSDO5 corpus. In general, our approach outperforms SOTA by a large margin on all domains and according to all evaluation metrics, especially when it comes to recall. We achieve results roughly twice as high as SOTA approach in F_1-score for all test domains regarding both monolingual and multilingual learning. One should note that the SOTA method was primarily meant for extracting terms from Ph.D. theses, i.e., documents significantly longer than those available in our training data, which explains the low

recall of that approach. However, this result clearly identifies a significant strength of the sequence-labeling approach - it does not rely on the frequency of term occurrences, which makes the approach more robust.

4.3.3 Relevance of the developed approaches

Given the rapid advancements in recent years, some of these approaches have already become outdated. Large language models (LLMs) represent a natural next step, with early work by H. T. H. Tran et al., 2024 showing promising results. Nonetheless, there remains room for improvement, although it is worth noting that this study utilizes a previous generation of LLMs; leveraging the latest models may yield significantly different outcomes.

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2022

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2022

A Transformer-based Sequence-labeling Approach to the Slovenian Cross-domain Automatic Term Extraction

Thi Hong Hanh Tran*†, Matej Martinc†, Andraž Repar†, Antoine Doucet‡, Senja Pollak†

*Jožef Stefan International Postgraduate School,
Jamova cesta 39, 1000 Ljubljana, Slovenia

† Jožef Stefan Institute,
Jamova cesta 39, 1000 Ljubljana, Slovenia

‡ University of La Rochelle,
23 Av. Albert Einstein, La Rochelle, France

Abstract

Automatic term extraction (ATE) is a popular research task that eases the time and effort of manually identifying terms from domain-specific corpora by providing a list of candidate terms. In this paper, we treat terminology extraction as a sequence-labeling task and experiment with a Transformer-based model XLM-RoBERTa to evaluate the performance of multilingual pretrained language models in the cross-domain sequence-labeling setting. The experiments are conducted on the RSDO5 corpus, a Slovenian dataset containing texts from four domains, including Biomechanics, Chemistry, Veterinary, and Linguistics. We show that our approach outperforms the Slovene state-of-the-art approach, achieving significant improvements in F1-score up to 40 percentage points. This indicates that applying multilingual pretrained language models for ATE in less-resourced European languages is a promising direction for further development. Our code is publicly available at https://github.com/honghanhh/sdjt-ate.

1. Introduction

Terms are single- or multi-word expressions denoting concepts from specific subject fields whose meaning may differ from the same set of words in other contexts or everyday language. They represent units of knowledge in a specific field of expertise and term extraction is useful for several terminographical tasks performed by linguists (e.g., construction of specialized term dictionaries). Most of these tasks are time- and labor-demanding, therefore recently several automatic term extraction approaches have been proposed to speed up the process.

Term extraction can also support and improve several complex downstream natural language processing (NLP) tasks. The broad range of downstream NLP tasks to which term extraction could benefit include, for example, glossary construction (Maldonado and Lewis, 2016), topic detection (El-Kishky et al., 2014), machine translation (Wolf et al., 2011), text summarization (Litvak and Last, 2008), information retrieval (Lingpeng et al., 2005), ontology engineering and learning (Biemann and Mehler, 2014), business intelligence retrieval (Saggion et al., 2007; Palomino et al., 2013), knowledge visualization (Blei and Lafferty, 2009), specialized dictionary creation (Le Serrec et al., 2010), sentiment analysis (Pavlopoulos and Androutsopoulos, 2014), and cold-start knowledge base population (Ellis et al., 2015), to cite a few.

In the attempt to ease the time and effort needed to manually identify terms from domain-specific corpora, automatic term extraction (ATE), also known as automatic term recognition (Kageura and Umino, 1996) or automatic term detection (Castellví et al., 2001), thus became an essential NLP task. However, despite the importance of term extraction and the research attention paid to the task, identifying the correct terms remains a notoriously challenging problem with the following not yet solved hurdles. First, despite several different definitions to describe the meaning of a term, the explicit distinction between terms and common words is in many cases still unclear. In addition, the characteristics of specific terms can vary significantly across domains and languages. Furthermore, the gold standard term lists and manually labeled domain-specific corpora for training and evaluation of ATE approaches are generally scarce for less-resourced languages including Slovenian, due to the large amount of work required for the construction of these resources.

Deep neural approaches towards ATE have been only recently proposed, but their evaluation in less-resourced languages has not yet been sufficiently explored and remains a research gap worth investigating. Inspired by the success of Transformer-based models in ATE from the recent TermEval 2020 competition's ACTER dataset (Hazem et al., 2020; Lang et al., 2021), we propose to exploit and explore the performance of XLM-RoBERTa pretrained language model (Conneau et al., 2019), which addresses the ATE as a sequence-labeling task. Sequence-labeling approaches have been successfully applied to a range of NLP tasks, including Named Entity Recognition (Lample et al., 2016; Tran et al., 2021) and Keyword Extraction (Martinc et al., 2021; Koloski et al., 2022). The experiments are conducted in the cross-domain setting on the RSDO5 corpus[1] (Jemec Tomazin et al., 2021a) containing Slovenian texts

[1] http://hdl.handle.net/11356/1470

from four domains (Biomechanics, Chemistry, Veterinary, and Linguistics).

The main contributions of this paper can be summarized in the following points:

- We systematically evaluate the performance of the Transformer-based pretrained model, namely XLM-RoBERTa, on the term extraction task, formulated as a supervised cross-domain sequence-labeling on the RSDO5 dataset containing texts from four different domains.
- We demonstrate that the proposed cross-domain approach surpasses the performance of the current state of the art (Ljubešić et al., 2019) for all the combinations of training and testing domains we experimented with, therefore establishing a new state-of-the-art (SOTA) method for the ATE on Slovenian corpus.

This paper is organized as follows: Section 2. presents the related work in the field of term extraction. Next, we introduce our methodology in Section 3., and the experimental details in Section 4.. The results with further error analysis are discussed in Section 5. and 6., before we conclude and present future works in Section 7..

2. Related Work

The history of ATE has its beginnings during the 1990s with research done by Damerau (1990), Ananiadou (1994), Justeson and Katz (1995), Kageura and Umino (1996), and Frantzi et al. (1998). ATE systems usually employ the following two-step procedure: (1) extracting a list of candidate terms; and (2) determining which of these candidate terms are correct using supervised or unsupervised approaches. Recently, neural approaches have been proposed.

Traditionally, the approaches were strongly based on linguistic knowledge and distinctive linguistic aspects of terms in order to extract possible candidates. Several NLP tools, such as tokenization, lemmatization, stemming, chunking, PoS tagging, full syntactic parsing, etc., are employed in this approach to obtain linguistic profiles of term candidates. As a heavily language-dependent approach, the better the quality of the pre-processing tools (e.g., FLAIR (Akbik et al., 2019), Stanza (Qi et al., 2020)), the better the quality of linguistic ATE methods.

Meanwhile, several studies preferred the statistical approach or combined linguistic and statistical approaches. Some of the measures include the termhood (Vintar, 2010), unithood (Daille et al., 1994) or C-value (Frantzi et al., 1998). Many current systems still apply some variation of this approach, most commonly in hybrid systems combining linguistic and statistical information (Repar et al., 2019; Meyers et al., 2018; Drouin, 2003; Macken et al., 2013; Šajatović et al., 2019; Kessler et al., 2019, to cite a few.).

Recently, advances in embeddings and deep neural networks have also influenced the term extraction field. Several embeddings have been investigated for term extraction, for example, uni-gram term representations constructed from a combination of local and global vectors (Amjadian et al., 2016), non-contextual word embeddings (Wang et al., 2016; Khan et al., 2016; Zhang et al., 2017), contextual word embeddings (Kucza et al., 2018), and the combination of both representations (Gao and Yuan, 2019).

In the recent ATE challenge, namely TermEval 2020 (Rigouts Terryn et al., 2020), the use of language models became very important. The winning approach on the Dutch corpus used pretrained GloVe word embeddings fed into a bi-directional LSTM based neural architecture. Meanwhile, the winning approach on the English corpus (Hazem et al., 2020) relied on the extraction of all possible n-gram combinations, which are fed into a BERT binary classifier that determines for each n-gram inside a sentence, whether it is a term or not. Besides BERT, several other variations of Transformer-based models have also been investigated. For example, RoBERTa and CamemBERT have been used in the TermEval 2020 challenge (Hazem et al., 2020). Another recent method is the HAMLET system (Rigouts Terryn et al., 2021), which proposes a hybrid adaptable machine learning approach that combines the linguistic and statistical clues to detect terms and is also evaluated on the TermEval data.

Meanwhile, Conneau et al. (2019) and Lang et al. (2021) take advantage of XLM-RoBERTa (XLM-R) to compare three different approaches, including a binary sequence classifier, a sequence classifier, and a token classifier employing the sequence-labeling approach (also under research by Kucza et al. (2018)), as we do in our research. Finally, Lang et al. (2021) proposes to use a multilingual encoder-decoder model called mBART (Liu et al., 2020), which is based on denoising pre-training, that generates sequences of comma-separated terms from the input sentences.

Annotated Corpora for Term Extraction Research (ACTER) dataset was released for the TermEval competition as a collection of four domain-specific corpora (Corruption, Wind energy, Equitation, and Heart failure) in three languages (English, French, and Dutch). However, when it comes to ATE for less-resourced languages, there is still a lack of gold standard corpora and limited use of neural methods. In recent years, the Slovene KAS corpus was compiled (Erjavec et al., 2021), and most recently the RSDO corpus that we use in our study (Jemec Tomazin et al., 2021b). Regarding the Slovenian language on which we focus in our study, the current SOTA was proposed by Ljubešić et al. (2019) that extracts the initial candidate terms using the CollTerm tool (Pinnis et al., 2019), a rule-based system employing a complex language-specific set of term patterns (e.g., POS tag,...) from the Slovenian SketchEngine module (Fišer et al., 2016), followed by a machine learning classification approach with features representing statistical term extraction measures. Another recent approach by (Repar et al., 2019) focuses on term extraction and alignment, where the main novelty is in using an evolutionary algorithm for the alignment of terms. On the other hand, the deep neural approaches have not been explored for Slovenian yet. Another problem very specific for less-resourced languages is that the open-sourced code is often not available for most current benchmark systems, hindering their reproducibility (for Slovenian, only the code by Ljubešić et al. (2019) is available).

Figure 1: An example of the (B-I-O) mechanism on a text sequence from Slovenian corpus.

3. Methodology

We consider ATE as a sequence-labeling task where the model returns a label for each token in a text sequence. We use the (B-I-O) labeling mechanism (Rigouts Terryn et al., 2021; Lang et al., 2021) where B stands for the beginning word in the term, I stands for the word inside the term, and O stands for the word not part of the term. The terms from a gold standard list are first mapped to the tokens in the raw text and each word inside the text sequence is annotated with one of the three labels (see examples in Figure 1). The model is first trained to predict a label for each token in the input text sequence (e.g., we model the task as token classification) and then applied to the unseen text (test data). Finally, from the tokens or token sequences labeled as terms, the final candidate term list for the test data is composed.

We experiment with XLM-RoBERTa [2] (Conneau et al., 2019), a Transformer-based model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. With the proliferation of non-English models (e.g., CamemBERT for French, Finnish BERT, German BERT, etc), XLM-RoBERTa, the multilingual version of RoBERTa (Liu et al., 2019), is a generic cross-lingual sentence encoder that achieves benchmark performance on multiple downstream NLP tasks, including ATE for rich-resourced languages (e.g. English) (Rigouts Terryn et al., 2020). Due to this well-documented SOTA performance on several related tasks, we opted to employ XLM-RoBERTa in a monolingual setting on our low-resourced Slovenian corpus. The overall architecture of our approach is presented in Figure 2.

In our experiments, we use a multilingual pre-trained language model in order to leverage the general knowledge the model obtained during pretraining on the huge multilingual corpus. First, we divide the dataset into train-validation-test splits. We also investigate the effectiveness of cross-domain learning, where the main idea is to test the transfer of knowledge from one domain to another and therefore evaluate the capability of the model to extract terms in new unseen domains as well as the ability to learn the relations between terms across domains given the assumption that they have terminologically-marked contexts. Therefore, we fine-tune the model on two domains (e.g., Biomechanics, Chemistry) as the train split, validate on a third domain (e.g., Veterinary) as the validation split, and test on the fourth domain that does not appear in the train set (e.g., Linguistics). The train split is used for fine-tuning the pre-trained language model. The validation split is applied to prevent over-fitting during the fine-tuning phase. Finally, the test split, which is not adopted during training, is used for the evaluation of the method.

[2] https://huggingface.co/xlm-roberta-base

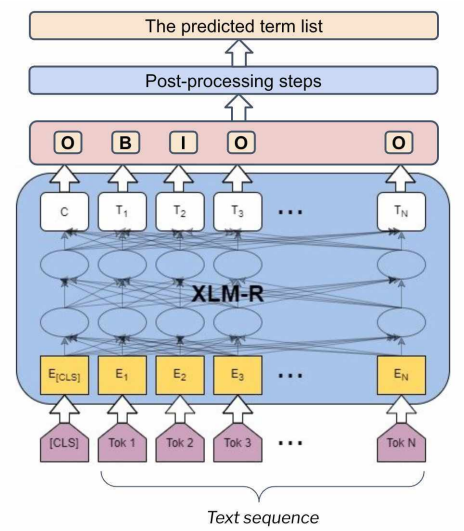

Figure 2: The overall architecture.

The model is fine-tuned on the training set to predict the probability for each word in a word sequence whether it is a part of the term (B, I) or not (O). To do so, an additional token classification head containing a feed-forward layer with a softmax activation is added on top of the model.

4. Experimental Setup

Here, we describe the dataset, the experimental details, and the metrics that we apply for the evaluation.

4.1. Dataset

The experiments are conducted on the Slovenian RSDO5 corpus version 1.1 (Jemec Tomazin et al., 2021a), which is a less-resourced Slavic language with rich morphology. As a part of the RSDO national project, the RSDO5 corpus was manually compiled and annotated and contains 12 documents with altogether about 250,000 words from the fields of Biomechanics (bim), Chemistry (kem), Veterinary (vet), and Linguistics (ling). The data were collected from diverse sources, including Ph.D. theses (3), a Ph.D. thesis-based scientific book (1), graduate-level textbooks (4), and journal articles (4) published between 2000 and 2019. Apart from the manually annotated terms, RSDO5 is also annotated with Universal Dependency tags (e.g. tags annotating tokens, sentences, lemmas, morphological features, etc.). However, in our research, we only leverage the original text with the term labels, where we consider all terms and do not distinguish between in-domain and out-of-domain terms.

In Table 1, we report on the number of documents, tokens, and unique terms across domains. Given the same

Languages	Biomechanics (bim)			Chemistry (kem)			Veterinary (vet)			Linguistics (ling)		
	# Docs	# Tokens	# Terms	# Docs	# Tokens	# Terms	# Docs	# Tokens	# Terms	# Docs	# Tokens	# Terms
Slovenian	3	61,344	2,319	3	65,012	2,409	3	75182	4,748	3	109,050	4,601

Table 1: Number of documents, tokens, and unique terms per domain in Slovenian RSDO5 dataset.

Languages	Biomechanics (bim)				Chemistry (kem)				Veterinary (vet)				Linguistics (ling)			
	B	I	O	% Term	B	I	O	% Term	B	I	O	% Term	B	I	O	% Term
Slovenian	7,070	6,835	47,439	22.67	7,614	4,486	52,912	18.61	10,953	6,261	57,968	22.90	12,348	6,079	90,623	16.89

Table 2: Label distribution and the proportion of terms appearing per domain in the Slovenian RSDO5 dataset.

number of collected documents for each domain, the documents from the Linguistics and Veterinary domains are longer (i.e., have more tokens) and also contain more terms than the domains of Biomechanics and Chemistry. In addition, Figure 3 presents the frequency of terms of different lengths per domain. Veterinary, Chemistry, and Linguistics share a similar term length distribution with most terms made of one to three words and only a few (less than three) terms longer than seven words (an example of a long term found in the corpus would be *"kaznivo dejanje zoper življenje , telo in premoženje"*, which means a crime against life, body, and property). Meanwhile, the Biomechanics domain distribution has a longer right tail, containing several terms with more than three words.

Furthermore, the corpus contains several nested terms, i.e., they also appear within larger terms and vice versa, a multiword term may contain shorter terms. For example, in the Biomechanics domain, term *"navor"* (torque) appears in terms such as *"sunek navora"* (torque shock), *"zunanji sunek navora"* (external torque shock), and *"izokinetični navor"* (isokinetic torque), to mention a few. This makes the labeling harder and the classifier needs to infer from the context whether a specific term is part of a longer term.

4.2. Implementation Details

We experiment with several combinations of training, validation, and testing data where two domains are used for training, the third one for validation, and the fourth one for testing (i.e., we train 12 models covering all possible domain combinations). We consider term extraction as a sequence-labeling or token classification task with a (B-I-O) annotation scheme. Table 2 presents the distribution across label types and the proportion of (B) and (I) labels in the total number of tokens per domain in the dataset. On average, the number of tokens annotated as terms (or parts of the term) only represents about one-fifth of the total tokens in the corpus, which means that there is a significant imbalance between (B) and (I) tokens, and tokens labeled as not terms (O).

We employ the XLM-RoBERTa token classification model and its "fast" XLM-RoBERTa tokenizer from the Huggingface library [3]. We fine-tune the model for up to 20 epochs regarding model convergence (i.e., we also employ the early stopping regime) with the learning rate of 2e-05, training and evaluation batch size of 32, and sequence length of 512 tokens, since this hyperparameter configuration performed the best on the validation set. The documents are split into sentences and the sentences containing more than 512 tokens are truncated, while the sentences with less than 512 tokens are padded with a special $< PAD >$ token at the end. During fine-tuning, the model is evaluated on the validation set after each training epoch, and the best-performing model is applied to the test set.

The model predicts each word in a word sequence whether it is a part of a term (B, I) or not (O). The sequences identified as terms are extracted from the text and put into a set of all predicted candidate terms. A post-processing step to lowercase all the candidate terms is applied before we compare our derived candidate list with the gold standard using the evaluation metrics discussed in Section 4.3..

4.3. Evaluation Metrics

We perform the global evaluation on our term extraction system by comparing the list of candidate terms extracted on the level of the whole test set with the manually annotated gold standard in the test set using Precision, Recall, and F1-score. Precision refers to the percentage of the extracted terms that are correct. Meanwhile, Recall indicates the percentage of the total correct terms that are extracted. Low Precision means a lot of noise in extraction whereas low Recall indicates the presence of lots of misses in extraction. Besides, F_1-score is a measure that computes an overall performance by calculating the harmonic mean between Precision and Recall). These evaluation metrics have been used also in the related work, including the TermEval 2020 shared task (Hazem et al., 2020; Rigouts Terryn et al., 2020; Lang et al., 2021).

5. Results

Table 3 presents the results achieved by the multilingual XLM-RoBERTa pre-trained language model on the Slovenian RSDO5 dataset. Note that the results in the table are grouped according to the model's test domain for better comparison between different settings. Our cross-domain approach proves to have relatively consistent performance across all the combinations, achieving Precision of more than 62%, Recall of no less than 55%, and F1-score above 61%. The model performs slightly better for the Linguistics and Veterinary domains than for Biomechanics and Chemistry. The difference in the number of terms and length of terms per domain pointed out in Section 4.1. might be one of the factors that contribute to this behavior. In addition, a significant performance boost can be observed for the Linguistics domain when the model is trained in the Chemistry

[3] https://huggingface.co/models

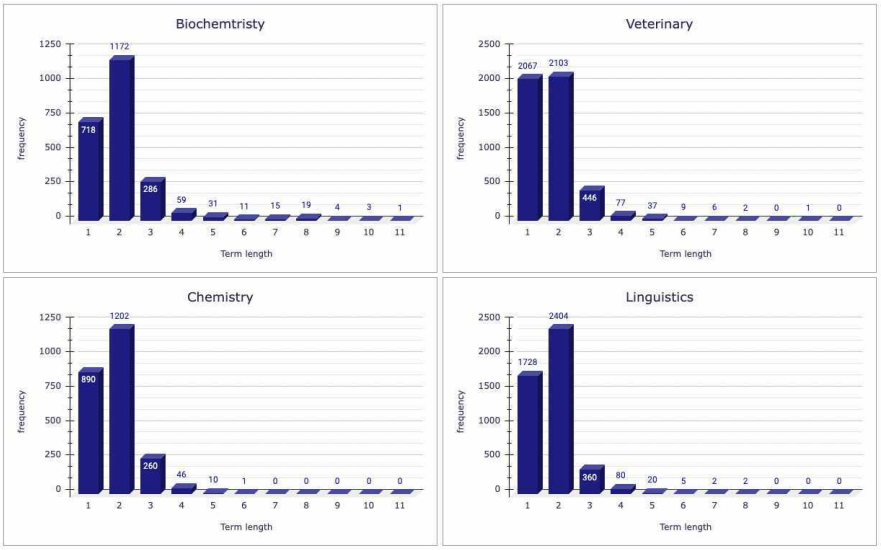

Figure 3: The frequencies of terms of specific length per each domain in a Slovenian dataset.

and Veterinary domains, and for the Veterinary domain, when the model is trained in Biomechanics and Linguistics. In these two settings, the model achieves an F1-score of more than 68%.

Training	Validation	Testing	Precision	Recall	F1-score
bim + kem	vet	ling	**69.55**	64.05	66.69
bim + vet	kem	ling	69.48	**73.66**	**71.51**
kem + vet	bim	ling	66.20	72.38	69.15
Ljubešić et al. (2019)		ling	52.20	25.40	34.10
bim + kem	ling	vet	71.06	66.72	68.82
bim + ling	kem	vet	**72.66**	65.59	**68.94**
ling + kem	bim	vet	69.3	**68.07**	68.68
Ljubešić et al. (2019)		vet	66.90	19.30	29.90
bim + vet	ling	kem	68.67	55.13	61.16
bim + ling	vet	kem	70.14	**60.27**	**64.83**
ling + vet	bim	kem	**70.23**	59.24	64.27
Ljubešić et al. (2019)		kem	47.80	31.40	37.80
vet + kem	ling	bim	**63.51**	**66.80**	**65.11**
vet + ling	kem	bim	62.25	65.20	63.69
ling + kem	vet	bim	62.35	63.99	63.16
Ljubešić et al. (2019)		bim	53.80	24.80	33.90

Table 3: Term extraction evaluation in a cross-domain setting on a Slovenian RSDO5 dataset.

We also present results for the current SOTA approach from Ljubešić et al. (2019) by reproducing their methodology in the same RSDO5 dataset. In general, our approach outperforms the approach proposed by Ljubešić et al. (2019) by a large margin on all domains and according to all evaluation metrics. The margin is especially large when it comes to Recall. Given the training process applied on RSDO5 corpus, Ljubešić et al. (2019) approach has low performance in F1-score due to the high imbalance between the Precision and Recall. This is most likely due to the fact that the methods employed by Ljubešić et al. (2019) rely heavily on the frequency and are thus not suitable for discovering low-frequency terms of which there are a lot in the RSDO5 corpus. In their own experiments, Ljubešić et al. (2019) discard all term candidates with a frequency below 3, hence why their results on their corpus are higher than on RSDO5.

Overall, we achieve results roughly twice as high as the approach proposed by Ljubešić et al. (2019) in terms of F1-score for all test domains. The results demonstrate the predictive power of contextual information in language models such as XLM-RoBERTa over the machine learning approach with features representing statistical term extraction measures as in Ljubešić et al. (2019).

6. Error Analysis

In this section, we analyze the predictions of XLM-RoBERTa in the RSDO5 corpus to get a better understanding of the model's performance and discover possible avenues for future work. First, we analyze the predictive power of our approach for terms of different lengths by calculating the Precision and Recall separately for terms of length k = {1,2,3,4, equal or more than 5}. The number of predicted candidate terms, number of ground truth terms, number of correct predictions (TPs), Precision, and Recall regarding different terms of length k and different test domains are presented in Tables 4, 5, 6, and 7. Note that these statistics are collected for the train-validation-test combinations that perform the best on each domain according to the F1-score.

Results across Tables 4 to 7 show that our models are good at predicting short terms containing up to three words in all four domains. The best model applied to the Linguistics test domain also shows competitive performance for the prediction of longer terms, achieving 75.00% Precision and a decent 31.03% Recall for terms with at least 5 words. Despite the relatively high Precision achieved by the models on long terms in the Veterinary and Biomechanics test domains, the Recall is pretty low, most likely due to the small

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	2,078	1,728	1,300	62.56	75.23
2	2,631	2,404	1,858	70.62	77.29
3	322	360	7,191	59.32	53.06
4	57	80	31	54.39	38.75
≥5	12	29	79	75.00	31.03

Table 4: Performance in Precision and Recall per term length in Linguistics domain.

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	2,159	2,067	1,472	68.18	71.21
2	2,062	2,103	1,448	70.22	68.85
3	314	446	182	57.96	40.81
4	28	77	10	35.71	12.99
≥5	3	55	2	66.67	3.64

Table 5: Performance in Precision and Recall per term length in Veterinary domain.

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	943	890	580	61.51	65.17
2	1,073	1,202	768	71.58	63.89
3	164	260	93	56.71	35.77
4	26	46	11	42.31	23.91
≥5	3	11	0	0.00	0.00

Table 6: Performance in Precision and Recall per term length in Chemistry domain.

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	1,079	718	22	48.38	72.70
2	1,153	1,172	822	71.29	70.14
3	223	286	124	55.61	43.36
4	26	59	11	42.31	18.64
≥5	11	84	5	45.45	5.95

Table 7: Performance in Precision and Recall per term length in Biomechanics domain.

amount of longer terms in the dataset on which the models are trained. When it comes to predictions in the Chemistry domain, there are no correct term predictions that consist of more than five words.

In addition, as the corpus contains many nested terms, the very common mistake the model makes is to predict a shorter term nested in the correct term of the gold standard (Pattern 1). Vice versa, the model sometimes generates incorrect predictions containing the correct nested terms (Pattern 2). Furthermore, in some cases, the model predicts a single prediction made out of two consecutive terms (Pattern 3). We report some examples of these incorrect patterns in Table 8, where the first column refers to the pattern type, the second one refers to our predicted candidate term, and the last column presents the true term from the gold standard. The presented candidate terms are extracted from the final list of predicted terms for the Linguistics test domain.

7. Conclusion

In summary, we investigated the performance of the multilingual Transformer-based language model, XLM-RoBERTa, in the monolingual cross-domain sequence-labeling term extraction task. The experiments were conducted on the representative Slovenian RSDO5 corpus, which contains texts from four specific domains, namely Biomechanics, Chemistry, Veterinary, and Linguistics. Our cross-domain sequence-labeling approach with XLM-RoBERTa had consistent performance across all the combinations of training, validation, and test set, achieving the performance of up to 72.66% in terms of Precision, up to 73.66% in terms of Recall, and up to 71.51% in terms of F1-score. The model performed slightly better in extracting terms from the Linguistics and Veterinary domains than from Biomechanics and Chemistry. Moreover, our approach outperformed the current state of the art on the Slovenian language (Ljubešić et al., 2019) by a large margin according to all three evaluation metrics, in some cases achieving three times higher Recall and roughly two times higher F1-score. As a consequence, our approach is the new SOTA approach on the RSDO5 dataset.

However, we believe that there remains room for improvement in the field of supervised term extraction. In the future, we would like to pre-train the model on the intermediate task (e.g., machine translation) resembling term extraction before fine-tuning it on the target downstream task, in order to boost the extraction performance. In addition, we will also investigate the performance of the models in the zero-shot cross-lingual setting, multi-lingual setting, and the combination of both settings in comparison with our current monolingual setting. Lastly, we suggest the integration of active learning into our current approach to improve the output of the automated method by dynamical adaptation after human feedback. By learning with humans in the loop, we aim at getting the most information with the least amount of term labels. We will also evaluate the contribution of active learning in reducing the annotation effort and determine the robustness of the incremental active learning framework across different languages and domains.

8. Acknowledgements

The work was partially supported by the Slovenian Research Agency (ARRS) core research program Knowledge Technologies (P2-0103) and project TermFrame (J6-9372), as well as the Ministry of Culture of the Republic of Slovenia through project Development of Slovene in Digital Environment (RSDO). The first author was partly funded by Region Nouvelle Aquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France.

9. References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair:

Patterns	Our predictions	The gold standards
1	*"klasična analogna telefonska zveza"* (classic analog telephone connection)	*"klasična analogna telefonska zveza pot"* (classic analog telephone connection path)
	"končnica neprve slovarske oblike" (suffix of non-first dictionary form)	*"končnica"* (suffix)
	...	...
2	*"brezžično slušalk v ušesu"* (wireless in-ear headphones)	*"brezžično slušalk"* (wireless headphones)
	"elektromehanska uporaba električne energije" (electromechanical use of electrical energy)	*"električne energije"* (electrical energy)
	...	...
3	*"batne parne stroje za pogon"* (reciprocating steam engines)	*"batne parne stroje" , "pogon"* (piston steam engines), (propulsion)
	"elektrarna na atomski pogon" (nuclear power plant)	*"elektrarna", "atomski pogon"* (power plant), (nuclear power plant)
	"besedilnim tipom strokovnega jezika" (text type professional language)	*"besedilnim tipom", "strokovnega jezika"* (text type), (professional language)
	"eksperimentalno modeliranje dinamičnih sistemov" (experimental modeling of dynamic systems)	*"eksperimentalno modeliranje", "dinamičnih sistemov"* (experimental modeling), (dynamic systems)
	...	...

Table 8: Examples of unlemmatised predictions in the Linguistics test domain.

An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Ehsan Amjadian, Diana Inkpen, Tahereh Paribakht, and Farahnaz Faez. 2016. Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 2–11.

Sophia Ananiadou. 1994. A methodology for automatic term recognition. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Chris Biemann and Alexander Mehler. 2014. *Text mining: From ontology learning to automated text processing applications*. Springer.

David M Blei and John D Lafferty. 2009. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

M Teresa Cabré Castellví, Rosa Estopa Bagot, and Jordi Vivaldi Palatresi. 2001. Automatic term detection: A review of current systems. *Recent advances in computational terminology*, 2:53–88.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Fred J Damerau. 1990. Evaluating computer-generated domain-oriented vocabularies. *Information processing & management*, 26(6):791–801.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *arXiv preprint arXiv:1406.6312*.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In *TAC*.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2021. The kas corpus of slovenian academic writing. *Language Resources and Evaluation*, 55(2):551–583.

Darja Fišer, Vit Suchomel, and Miloš Jakubícek. 2016. Terminology extraction for academic slovene using sketch engine. In *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, pages 135–141.

Katerina T Frantzi, Sophia Ananiadou, and Junichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *International conference on theory and practice of digital libraries*, pages 585–604. Springer.

Yuze Gao and Yu Yuan. 2019. Feature-less End-to-end Nested Term extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 607–616. Springer.

Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2020. TermEval 2020: TALN-LS2N System for Automatic Term Extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 95–100.

Mateja Jemec Tomazin, Mitja Trojar, Simon Atelšek, Tanja Fajfar, Tomaž Erjavec, and Mojca Žagar Karer. 2021a.

Corpus of term-annotated texts RSDO5 1.1. Slovenian language resource repository CLARIN.SI.

Mateja Jemec Tomazin, Mitja Trojar, Mojca Žagar, Simon Atelšek, Tanja Fajfar, and Tomaž Erjavec. 2021b. Corpus of term-annotated texts rsdo5 1.0.

John S Justeson and Slava M Katz. 1995. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural language engineering*, 1(1):9–27.

Kyo Kageura and Bin Umino. 1996. Methods of Automatic Term Recognition.A Review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.

Rémy Kessler, Nicolas Béchet, and Giuseppe Berio. 2019. Extraction of terminology in the field of construction. In *2019 First International Conference on Digital Data Processing (DDP)*, pages 22–26. IEEE.

Muhammad Tahir Khan, Yukun Ma, and Jung-jae Kim. 2016. Term Ranker: A Graph-Based Re-Ranking Approach. In *FLAIRS Conference*, pages 310–315.

Boshko Koloski, Senja Pollak, Blaž Škrlj, and Matej Martinc. 2022. Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? *arXiv preprint arXiv:2202.06650*.

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In *INTERSPEECH*, pages 2072–2076.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3607–3620.

Annaïch Le Serrec, Marie-Claude L'Homme, Patrick Drouin, and Olivier Kraif. 2010. Automating the compilation of specialized dictionaries: Use and analysis of term extraction and lexical alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(1):77–106.

Yang Lingpeng, Ji Donghong, Zhou Guodong, and Nie Yu. 2005. Improving retrieval effectiveness by using key terms in top retrieved documents. In *European Conference on Information Retrieval*, pages 169–184. Springer.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Coling 2008: Proceedings of the workshop multi-source multilingual information extraction and summarization*, pages 17–24.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. Kas-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. In *International Conference on Text, Speech, and Dialogue*, pages 115–126. Springer.

Lieve Macken, Els Lefever, and Veronique Hoste. 2013. Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1):1–30.

Alfredo Maldonado and David Lewis. 2016. Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In *Proceedings of The 12th International Conference on Terminology and Knowledge Engineering*, pages 91–100.

Matej Martinc, Blaž Škrlj, and Senja Pollak. 2021. Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, page 1–40.

Adam L Meyers, Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman, and Olga Babko-Malaya. 2018. The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores. *Frontiers in Research Metrics and Analytics*, 3:19.

Marco A Palomino, Tim Taylor, and Richard Owen. 2013. Evaluating business intelligence gathering techniques for horizon scanning applications. In *Mexican International Conference on Artificial Intelligence*, pages 350–361. Springer.

John Pavlopoulos and Ion Androutsopoulos. 2014. Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 44–52.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatjana Gornostaja, Špela Vintar, and Darja Fišer. 2019. Extracting data from comparable corpora. In *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*, pages 89–139. Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Andraž Repar, Vid Podpečan, Anže Vavpetič, Nada Lavrač, and Senja Pollak. 2019. TermEnsembler: An Ensemble Learning Approach to Bilingual Term Extraction and Alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 25(1):93–120.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared Task on

Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 85–94. European Language Resources Association (ELRA).

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2021. HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology*.

Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. 2007. Ontology-based information extraction for business intelligence. In *The Semantic Web*, pages 843–856. Springer.

Antonio Šajatović, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. 2019. Evaluating automatic term extraction methods on individual documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 149–154.

Thi Hong Hanh Tran, Antoine Doucet, Nicolas Sidere, Jose G Moreno, and Senja Pollak. 2021. Named entity recognition architecture combining contextual and global features. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings*, page 264. Springer Nature.

Spela Vintar. 2010. Bilingual Term Recognition Revisited: The Bag-of-equivalents Term Alignment Approach and its Evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(2):141–158.

Rui Wang, Wei Liu, and Chris McDonald. 2016. Featureless Domain-Specific Term Extraction with Minimal Labelled Data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 103–112.

Petra Wolf, Ulrike Bernardi, Christian Federmann, and Sabine Hunsicker. 2011. From statistical term extraction to hybrid machine translation. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Ziqi Zhang, Jie Gao, and Fabio Ciravegna. 2017. SemRe-Rank: Incorporating Semantic Relatedness to Improve Automatic Term Extraction Using Personalized PageRank. *arXiv preprint arXiv:1711.03373*.

Machine Learning (2024) 113:4285–4314
https://doi.org/10.1007/s10994-023-06506-7

Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling?

Hanh Thi Hong Tran[1,2,3] · Matej Martinc[2] · Andraz Repar[2] · Nikola Ljubešić[2] · Antoine Doucet[3] · Senja Pollak[2]

Received: 5 March 2023 / Revised: 12 August 2023 / Accepted: 16 December 2023 /
Published online: 27 March 2024

Abstract
Automatic term extraction (ATE) is a natural language processing task that eases the effort of manually identifying terms from domain-specific corpora by providing a list of candidate terms. In this paper, we treat ATE as a sequence-labeling task and explore the efficacy of XLMR in evaluating cross-lingual and multilingual learning against monolingual learning in the cross-domain ATE context. Additionally, we introduce NOBI, a novel annotation mechanism enabling the labeling of single-word nested terms. Our experiments are conducted on the ACTER corpus, encompassing four domains and three languages (English, French, and Dutch), as well as the RSDO5 Slovenian corpus, encompassing four additional domains. Results indicate that cross-lingual and multilingual models outperform monolingual settings, showcasing improved F1-scores for all languages within the ACTER dataset. When incorporating an additional Slovenian corpus into the training set, the multilingual model exhibits superior performance compared to state-of-the-art approaches in specific scenarios. Moreover, the newly introduced NOBI labeling mechanism enhances the classifier's capacity to extract short nested terms significantly, leading to substantial improvements in Recall for the ACTER dataset and consequentially boosting the overall F1-score performance.

Keywords Term extraction · XLMR · Sequence labeling · Cross-lingual · Cross-domain · Nested terms

Editors: Dino Ienco, Roberto Interdonato, Pascal Poncelet.

✉ Senja Pollak
senja.pollak@ijs.si

Hanh Thi Hong Tran
tran.hanh@ijs.si

[1] Jožef Stefan International Postgraduate School, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[2] Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[3] University of La Rochelle, 23 Av. Albert Einstein, La Rochelle, France

1 Introduction

Terms are textual expressions that denote concepts in a specific field of expertise. They are beneficial for several terminographical tasks performed by linguists (e.g., construction of specialized terminological dictionaries (Le Serrec et al., 2010)). Moreover, terms can also support and improve several downstream natural language processing (NLP) tasks (e.g., topic detection (ElKishky et al., 2014), information retrieval (Lingpeng et al., 2005), machine translation (Wolf et al., 2011)). To ease the time and effort needed to manually identify terms in domain-specific corpora, automatic term extraction (ATE) approaches were proposed.

The TermEval 2020 shared task, organized as part of the CompuTerm workshop (Rigouts et al., 2020a), presented one of the first opportunities to systematically study and compare several ATE architectures with the introduction of the Annotated Corpora for Term Extraction Research (ACTER) dataset (Rigouts et al., 2020a, b). While the workshop was a significant step forward in systematic comparison, the less-resourced languages (e.g., Slovenian) have not yet been sufficiently explored and remain a research gap. Furthermore, there is still room for improvement in performance. In our previous study (Tran et al., 2022a), the conducted error analysis pointed out that the two most common errors that the tested classifiers made were to predict a shorter term nested in the ground truth term and vice versa, i.e., the model sometimes generates the terms not covered in the ground truth, containing a nested term. This insight leads to a hypothesis about the insufficiency of the widely used BIO labeling regime (Hazem et al., 2020). This regime does not allow labeling the nested terms and giving the model the necessary information to avoid the above mistakes.

Inspired by the success of Transformers (Hazem et al., 2020) and the rise of cross-lingual learning (Lang et al., 2021), our research delves into the effectiveness of the XLMR (Conneau et al., 2020) in multilingual and cross-lingual scenarios. First, having a single model that works across several languages is important, as it can be used also in the languages not seen during the training. Instead of having to construct language-specific models, multilingual and cross-lingual models can be directly used on any new language that is supported by XLMR. In addition, for the languages where the data is available, having a single model instead of many language-specific models is a much simpler solution, and can also make the models less dataset-specific.

Our approach frames the ATE task as a sequence-labeling problem, as this strategy has proven successful in various NLP tasks like Named Entity Recognition (NER) (Lample et al., 2016; Tran et al., 2021) and Keyword Extraction (Martinc et al., 2021). Additionally, we extend our previous work (Tran et al., 2022a) by introducing an innovative nested term labeling mechanism, incorporating two extra labels for single nested terms, and rigorously evaluating the model's performance in cross-lingual and multi-lingual settings. This comprehensive exploration showcases the power of a multilingual pretrained language model with cross-lingual and multi-lingual settings in capturing and understanding diverse linguistic nuances. The experiments are conducted in the cross-domain setting on the ACTER dataset[1] containing texts in four domains (Corruption, Wind energy, Equitation, and Heart failure) with three languages (English, French, and Dutch) and the RSDO5 corpus[2] (Jemec

[1] https://github.com/AylaRT/ACTER.

[2] https://www.clarin.si/repository/xmlui/handle/11356/1470.

Tomazin et al., 2021) containing Slovenian texts from four domains (Biomechanics, Chemistry, Veterinary, and Linguistics).

The main contributions of this paper can be summarized as follows:

– We propose a new NOBI annotation mechanism to better capture single nested terms. When a dataset contains a relevant proportion of nested terms, the new labeling regime improves the Recall of the models by a large margin, leading also to further improvements in the F1-score. This is also the main novelty compared to the shorter conference version (Tran et al., 2022a) of this paper.
– We systematically evaluate the performance of the XLMR on the cross-domain term extraction task in two datasets covering English, French, Dutch, and a less-resourced Slovenian in both standard BIO and the novel NOBI scheme.
– We compare the performance among cross-lingual, multilingual, and monolingual approaches to determine the general applicability of multilingual language models for sequence labeling in both rich- and less-resourced languages. The datasets using BIO and NOBI annotation regimes are both considered.

2 Related work

The history of ATE has its beginnings during the 1990s with research done by Damerau (1990) and Justeson and Katz (1995). ATE systems usually employ the two-step procedure: (1) extracting a list of candidate terms, and (2) determining which candidate terms are correct.

2.1 Approaches based on term characteristics

Traditional approaches relied on distinctive linguistic aspects of terms to extract possible candidates. Several NLP tools (e.g., tokenization, lemmatization, stemming, PoS tagging) are employed to obtain linguistic profiles of term candidates. As a heavily language-dependent approach, the better the quality of the pre-processing tools (e.g., FLAIR (Akbik et al., 2019), Stanza (Qi et al., 2020)), the better the quality of linguistic methods. More recent studies preferred the statistical approach, which commonly relies on the assumption that a higher candidate term frequency in a domain-specific corpus implies a higher likelihood that a candidate is an actual term. Some measures relying on this assumption include termhood (Vintar, 2010), unithood (Daille et al., 1994) or C-value (Frantzi et al., 1998). More popular statistical approaches also considered the frequency of the term internal words compared to the term frequency to identify rare terms and remove frequent words. Many current systems still apply this approach's variation, or hybrid mechanisms that combine linguistic and statistical information (Kessler et al., 2019; Repar et al., 2019).

2.2 Approaches based on machine learning and deep learning

Recent advances in representation learning and deep neural networks have also influenced term extraction. Several embedding techniques have been investigated for the task at hand, e.g., uni-gram (Amjadian et al., 2016), non-contextual (Zhang et al., 2018), contextual (Kucza et al., 2018) word embeddings, and the hybrid ones (Gao & Yuan, 2019). The first use of language models for the ATE task was in the TermEval 2020 (Rigouts et al., 2020a)

where the winning approach on the Dutch corpus used BiLSTM-based neural architecture with GloVe embeddings while the winning solution on the English corpus (Hazem et al., 2020) extracted all possible n-gram combinations, which are then fed into a BERT binary classifier that determines for each n-gram inside a sentence, whether it is a term. Besides, several Transformer-based variations have also been investigated (e.g., RoBERTa, CamemBERT (Hazem et al., 2020)). Further work includes HAMLET by Terryn et al., 2021, which proposes a hybrid adaptable machine learning classifier that combines linguistic and statistical clues to detect terms.

Recently, sequence-labeling and cross-lingual approaches toward ATE have been gaining traction. Kucza et al. (2018) was one of the first to model term extraction as a sequence-labeling task. Cross-lingual sequence labeling was, on the other hand, explored in Conneau et al. (2020), Lang et al. (2021), Hazem et al. (2022), and Tran et al. (2022a), who took advantage of XLMR, the model we also employ in this work. Lang et al. (2021) compared different cross-lingual approaches, including a sequence classifier, and a token classifier on this sequence-labeling task, and further proposed a sequence-to-sequence (seq2seq) approach, which used mBART (Liu et al., 2020) to generate sequences of comma-separated terms from the input. The results demonstrate the capability of multilingual models to outperform monolingual ones in some specific scenarios and the potential of cross-lingual learning.

Finally, in our conference paper (Tran et al., 2022a) that we extend in this journal paper, we leveraged the multilingual setup by fine-tuning the model using training datasets from several languages and then applying the model to their test sets, separately. By doing so, we examined whether adding more data from other languages to the training set that matches the target language in the testing set improves the model's predictive performance. After adding the Slovenian corpus into the ACTER training set, our multilingual model demonstrated a significant improvement in Recall across all test languages compared to the monolingual one.

2.3 Approaches for Slovenian term extraction

For Slovenian, the language used in our study, and for less-resourced languages in general, the research is still hindered by the lack of gold standard corpora and limited use of neural methods. Things are nevertheless slowly improving. In recent years, the Slovenian KAS corpus was compiled (Erjavec et al., 2021), quickly followed by another corpus designed for term extraction, the RSDO5 corpus.[3] Regarding the methods, Vintar (2010) was one of the first to propose statistical approaches for Slovenian ATE tasks. After that, Ljubešić et al. (2019) introduced a hybrid one, in which they extract the initial candidate terms using the CollTerm tool (Pinnis et al., 2019), a rule-based system employing a complex language-specific set of term patterns from the Slovenian SketchEngine (Fišer et al., 2016). Meanwhile, Repar et al. (2019) focuses on term extraction and alignment, where the novelty is the evolutionary algorithm for the term alignment.

The deep neural approaches have not been sufficiently explored for Slovenian data yet. The only neural approach towards Slovenian ATE was proposed in our recent study (Tran et al., 2022b). There, we implemented the Transformers-based sequence-labeling approach, which we extend in this study, in a cross-lingual and multilingual evaluation. Another

[3] https://www.clarin.si/repository/xmlui/handle/11356/1470.

problem is that often no open-sourced code is available for most current benchmark systems, hindering their reproducibility (for Slovenian, only the code from Ljubešić et al. (2019) and Tran et al. (2022b) methods are available).

2.4 Extraction of nested terms

In many practical applications, it is common that the terms have a nested structure where a term could contain other terms or be part of others. Vintar (2004) first suggested ranking and/or discarding nested terms using the C-value, but their results were unsatisfactory. Marciniak and Mykowiecka (2015) later identified them by combining grammatical correctness and normalized pointwise mutual information (NPMI) based on bigrams in a corpus. However, this method's efficiency relies heavily on corpus features (e.g., size, thematic homogeneity, and phrase frequency). Recently, Gao and Yuan (2019) proposed an end-to-end architecture that employs classification and ranking for n-gram candidates in text sequences. Nonetheless, this suffers from reduced Recall due to ranking and its threshold output is not applicable to new, unseen domains. Since then, no further methodologies have been proposed, leaving a gap in extracting nested terms for term extraction tasks.

Regarding other NLP downstream tasks sharing the same mechanisms (e.g., NER, Keyword Extraction), besides the common sequence tag schemes (e.g., BIO (Ramshaw & Marcus, 1999), IOBES (Lester, 2020), BMEWO (Ratinov & Roth, 2009), BILOU (Ratinov & Roth, 2009)) for both flat and nested ones, we can categorize the methods to capture nested entities into four main types: (1) sequence labeling, (2) hypergraph-based, (3) sequence-to-sequence (Seq2Seq), and (4) span-based methods. However, none of them except the BIO regime for the sequence-labeling approach has been applied for term extraction yet.

3 Methodology

Section 3.1 presents a brief description of our chosen datasets. We demonstrate the general methodology, experimental setup, and implementation details in Sects. 3.2 and 3.3. Finally, in Sect. 3.4, we present our chosen evaluation metrics.

3.1 Datasets

The experiments were conducted on ACTER (Rigouts et al., 2020a) and RSDO5 version 1.1 (Jemec Tomazin et al., 2021), both comprising texts from diverse languages and domains. The ACTER dataset is a collection of 12 corpora covering four domains (Corruption (corp), Equitation (equi), Wind energy (wind), and Heart failure (htfl)) in three languages (English (en), French (fr) and Dutch (nl)). The dataset has two types of gold standard annotations: one containing both terms and named entities (NES), and the other one containing only terms (ANN). The second dataset is the RSDO5 version 1.1 (Jemec Tomazin et al., 2021), which contains texts in Slovenian (sl), a less-resourced Slavic language with rich morphology. The corpus contains 12 documents collected from 2000 to 2019 covering domains of Biomechanics (bim), Chemistry (kem), Veterinary (vet), and Linguistics (ling). The data analysis is depicted in Figs. 14, 15 , 16, 17 and 18 and Table 7.

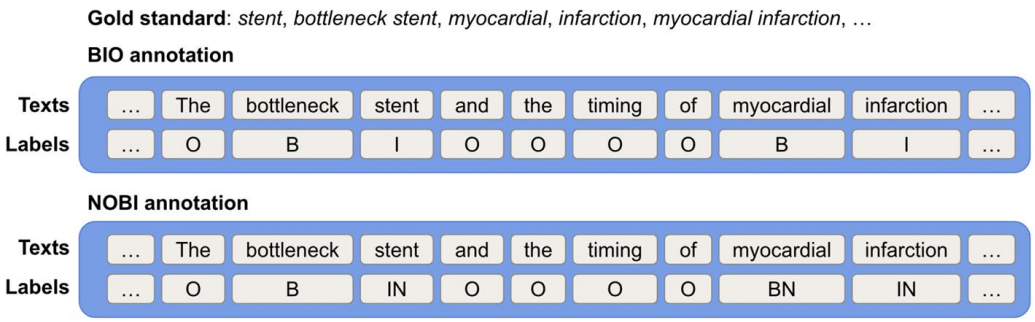

Fig. 1 An example of BIO and NOBI annotation regimes in the ACTER corpus

3.2 Experimental setup

We consider ATE as a sequence-labeling task where the model returns a label for each token in a text sequence using two different labeling regimes: the benchmark BIO labeling scheme (Lang et al., 2021; Rigouts et al., 2021) and our novel annotation scheme called NOBI. In the BIO regime, B stands for the beginning word in the term, I stands for the word inside the term, and O stands for the word not part of the term. The terms from a gold standard list are first mapped to the tokens in the raw text and each word inside the text sequence is annotated with one of three labels (see the upper example in Fig. 1). However, it is not optimized for nested term extraction. Thus, we propose NOBI, an annotation regime with two additional labels BN and IN, referring to a word being in the beginning or inside the nested term, respectively (see the lower example in Fig. 1). An annotation regime with two additional labels BN and IN, where N refers to nested single-word terms, which can be at the beginning (BN) or inside (IN) position of a longer term.

In Fig. 1, the gold standard contains the following terms: *"stent", "bottleneck stent", "myocardial", "infarction", "myocardial infarction"*, etc. In the BIO regime, we ignore the single nested terms, thus, we only mark *"bottleneck"* as the beginning (B) and *"stent"* as the inside (I) of the full term *"bottleneck stent"*. Similarly, *"myocardial"* is the beginning (B), and *"infarction"* is the inside (I) of the full term *"myocardial infarction"*. However, in the NOBI regime, we consider *"bottleneck stent"* and *"stent"* as two different terms where *"stent"* is the nested term of *"bottleneck stent"*, in contrast to the BIO scheme, where the model extracts just the *"bottleneck stent"* as a term. Similarly, *"myocardial"* and *"infarction"* are two separate terms that are nested in *"myocardial infarction"*. Therefore, an additional label N is added to the label of *"stent"*, *"myocardial"*, and *"infarction"*.

We do not consider either multi-word nested terms or terms nested in other nested terms – so-called nested terms on the second or higher levels – due to their rarity in the corpora and gold standards (see the nested frequency in the gold standard from Figs. 16 and 18 in Appendix). Despite the difference in the number of terms in each language and domain, the percentage of unique nested terms in all languages and domains is somewhat consistent, ranging around one-third of the total unique terms in the gold standards. However, the number of terms nested in other nested terms only takes one-tenth and one-twelfth of the total amount of unique terms in both corpora, respectively, and the amounts are even much smaller if we specify the ratio per nested level (e.g., in the second level, third level). We also demonstrate in Table 7 in Appendix the proportion of the nested terms with different

word lengths k where $k = \{1, 2, 3, 4, \geq 5\}$ for each domain and language of both corpora. The last column on the right calculates the percentage of single-word nested terms in total nested terms in the first level. On average, the amount of single-word nested terms accounts for 78.06% above all the nested terms on the first levels in the corpora. Therefore, we only label single-word nested terms on the first level.

For both labeling regimes, we experiment with XLMR, a Transformer-based model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. The model is first trained to predict a label for each token in the input text sequence (e.g., we model the task as token classification) and then applied to the unseen text (test data). Finally, from the tokens or token sequences labeled as terms, the final candidate term list for the test data is composed. Note that when the NOBI annotation regime is used, the terms labeled with BN and IN are added to the final term list separately, together with the terms in which they are nested.

We evaluate the cross-domain performance of XLMR in monolingual, cross-lingual, and multilingual settings. Altogether, 78 different scenarios per annotation regime are tested. The distinct settings are described below.

1. **Monolingual setup.** We evaluate how well the model performs when there is a language-specific training corpus available and there is a match between the language of the train set and the language of the test set. For better comparison with other existing approaches, we apply the same configuration as in the TermEval 2020 shared task where Heart failure of each language is considered as the test set. Thus, we fine-tune our model on a single language, which means we train three monolingual models for three languages (English, French, Dutch) and test each model in the same language for each annotation regime. Besides, we train 12 monolingual models for each annotation regime for Slovenian given 12 different combinations of train-validation-test split regarding the domains.
2. **Cross-lingual setup.** We evaluate the capability of the model to apply the knowledge learned in one or more languages for ATE in another unseen language. Therefore, we fine-tune the ATE model on one or more languages (e.g., English and Dutch) and test it on another language not appearing in the train set (e.g., French). In this scenario, we, therefore, examine how well the model performs without the language-specific training corpus and how good the knowledge transfer between different languages is.
3. **Multilingual setup.** We fine-tune our model using a.) training datasets from the languages in the ACTER dataset (English, French, and Dutch) or b.) training datasets from the languages in the ACTER dataset plus the Slovenian training dataset from the RSDO5 corpus, and then apply the model to the test sets of all languages in the ACTER dataset. By doing so, we examine whether adding more data from other languages to the training set in the target language improves the predictive performance of the model.

All three settings are applied in a cross-domain evaluation scenario, where we use two domains for training, another domain for validation, and the rest for testing. One exception is the multilingual and cross-lingual settings with the additional Slovenian corpus in the training set, where we use two domains from ACTER corpora and all domains from the RSDO5 corpus for the training. This way, we can evaluate the model's generalization capabilities to adapt knowledge in one or more domains to a new, unseen arbitrary one and, therefore, much more useful. In the ACTER dataset, we use the Corruption and Wind energy domains for training, the Equitation domain for validation, and the Heart

failure domain for testing, in order to allow for a direct comparison with other benchmark approaches from the related work, which employ the same train-validation-test setting (Lang et al., 2021). Meanwhile, in the RSDO5 corpus, we explore different train-validation-test combinations.

We divide the dataset into train-validation-test splits. The model is fine-tuned on the training set to predict the probability for each word in a word sequence whether it is a part of the term (B, I), whether it is a nested term (BN for nested terms at the beginning of a multi-word term, IN for nested terms at non-beginning positions of a multi-word term), or not part of the term (O). To do that, an additional token classification head containing a feed-forward layer with a softmax activation is added on top of each model.

3.3 Implementation details

We employ the XLMR token classifier from Huggingface.[4] We fine-tune the model for up to 20 epochs (i.e., the early stopping regime via the validation set) using the learning rate of 2e-05, training and evaluation batch size of 32, and sequence length of 512 tokens, since this hyperparameter configuration performed the best on the validation set. First, the documents are split into sentences. Then, the sentences with more than 512 tokens are truncated, while those with less than 512 tokens are padded with a special $< PAD >$ token at the end.

During fine-tuning, the model is evaluated on the validation set after each training epoch, and the best-performing model is applied to the test set. Note that the model with BIO annotation regime will predict the probability of whether the word is a part of the terms (B, I) or not (O) while the one with NOBI regime will predict the probability in the same manner as BIO except for the additional information on the nested terms of the first level (BN, IN) where each word with the N label will be considered as an individual single-word candidate term. The sequences identified as terms are extracted from the text and put into a set of all predicted candidate terms. A post-processing step to lowercase all candidate terms is applied before we compare our derived candidate list with the gold standard.

3.4 Evaluation metrics

We evaluate the performance of the ATE systems by comparing the candidate list extracted from the test set with the manually annotated gold standard term list for that specific test set. We use exact string matching to compare the retrieved terms to the ones in the gold standard and calculate Precision (P), Recall (R), and F1-score (F1). These evaluation metrics have also been used in related work (Hazem et al., 2020; Lang et al., 2021; Rigouts et al., 2020a; Ljubešić et al., 2019), therefore, our results are directly comparable to the benchmarks.

[4] https://huggingface.co/models.

Table 1 Evaluation on the ACTER dataset given Heart failure as a test set. For each test set, bold is used to indicate the best model in terms of P, R, and F1 for each test set (ANN and NES) and each annotation scheme separately (BIO and NOBI). The arrows are used for the comparison of BIO and NOBI for each setting, where ↑ is used to show the better performance of NOBI compared to BIO, while ↓ denotes the lower performance of NOBI compared to BIO. In blue, we indicate the best model in terms of the F1 for each test set

Train language	ANN						NES					
	BIO			NOBI			BIO			NOBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
English test set.												
en	58.1	48.1	52.6	↓ 57.5	↑ 48.6	↑ 52.7	62.1	52.1	56.7	↓ 58.6	↑ 55.2	↑ 56.9
fr	56.9	33.2	42.0	↓ 54.2	↑ 34.7	↑ 42.3	60.0	39.1	47.4	↓ 57.8	↑ 44.3	↑ 50.2
nl	55.6	56.4	**56.0**	↑ 57.6	↑ 58.4	↑ **58.0**	54.4	57.7	56.0	↑ 56.9	↑ 61.2	↑ 59.0
fr, sl	47.1	65.8	54.9	↓ 42.5	↑ 68.8	↓ 52.5	49.2	64.3	55.7	↓ 44.6	↑ 66.6	↓ 53.4
nl, sl	45.7	**66.3**	54.1	↑ 46.0	↑ 67.8	↑ 54.8	48.1	65.4	55.5	↑ 49.2	↑ 67.0	↑ 56.8
fr, nl	**60.8**	46.8	52.9	↓ 57.5	↓ 41.5	↓ 48.2	62.3	50.5	55.7	↓ 58.6	↑ 52.0	↑ 55.1
fr, nl, sl	50.0	62.4	55.5	↓ 48.3	↑ 67.2	↑ 56.2	52.1	63.2	57.2	↓ 49.5	↑ 65.3	↓. 56.3
en, fr	57.2	51.2	54.0	↑ **58.0**	51.2	↑ 54.4	60.4	51.5	55.6	↓ **59.5**	↑ 54.2	↑ 56.7
en, nl	58.0	48.7	52.9	↓ 54.0	↑ 56.1	↑ 55.0	**62.4**	51.4	56.4	↓ 57.4	↑ 58.6	↑ 58.0
en, sl	48.1	63.2	54.6	↑ 49.0	↑ 65.7	↑ 56.1	54.9	63.8	59.0	↓ 50.8	↑ 64.4	↓ 56.8
en, fr, sl	48.1	64.2	55.0	↑ 51.1	↑ 67.2	↑ **58.0**	58.4	61.1	**59.7**	↓ 55.2	↑ 63.4	↓ 59.0
en, nl, sl	48.4	65.0	55.4	↓ 44.8	↑ 68.6	↓ 54.2	54.5	63.3	58.6	↓ 53.1	↑ 67.3	↑ **59.3**
en, fr, nl	56.8	53.0	54.9	↓ 55.7	↓ 51.0	↓ 53.3	60.8	52.6	56.4	↓ 57.4	↑ 59.8	↑ 58.6
en, fr, nl, sl	45.9	**66.3**	54.2	↓ 45.5	↑ **69.3**	↑ 54.9	48.3	**65.7**	55.6	↑ 51.9	↑ **68.4**	↑ 59.0
cross-ling. avg.	52.7	55.2	52.6	↓ 51.0	↑ 56.4	↓ 52.0	54.4	56.7	54.6	↓ 52.8	↑ 59.4	↑ 55.1
multi-ling. avg.	51.8	58.8	54.4	↓ 51.2	↑ 61.3	↑ 55.1	57.1	58.5	57.3	↓ 55.0	↑ 62.3	↑ 58.2
French test set.												
fr	**70.5**	44.4	54.5	↓ 66.3	↑ 48.9	↑ 56.3	**72.4**	48.5	58.1	↓ 65.9	↓ 54.7	↑ 59.8
en	66.7	47.9	55.8	↑ 67.8	↓ 44.8	↓ 53.9	70.6	53.8	61.1	↓ 65.3	↓ 53.3	↓ 58.7
nl	66.5	51.5	58.0	↓ 64.7	↓ 47.0	↓ 55.1	67.6	53.2	59.5	↓ 67.5	↓ 53.1	↓ 59.4
en, sl	60.2	61.4	60.8	↑ 61.1	↓ 57.5	↓ 59.2	57.8	**62.5**	60.1	↑ 62.9	↓ 56.0	↓ 59.2
nl, sl	61.4	60.4	60.9	↓ 59.5	↓ 58.5	↓ 59.0	61.8	59.9	60.8	↑ 63.1	↓ 56.7	↓ 59.7
en, nl	65.3	44.2	52.7	↓ 65.2	↑ 47.9	↑ 55.2	68.7	52.4	59.4	↑ **69.3**	↓ 50.6	↓ 58.5
en, nl, sl	58.7	61.0	59.8	↓ 55.3	↑ 63.2	↓ 59.0	60.9	62.0	61.5	↓ 59.0	↑ 62.3	↓ 60.6
fr, en	63.7	52.4	57.5	↑ 65.7	↓ 49.5	↓ 56.4	68.1	52.8	59.5	↓ 67.2	↓ 49.6	↓ 57.1
fr, nl	69.2	48.3	56.9	↓ **66.4**	↑ 48.4	↓ 56.0	70.7	49.5	58.3	↓ 66.1	↑ 54.2	↑ 59.6
fr, sl	65.0	56.6	60.5	↓ 58.8	↑ 62.3	60.5	65.3	57.6	61.2	↓ 56.9	↑ 64.0	↓ 60.2
fr, en, sl	61.5	58.6	60.0	↑ 63.2	↑ 60.5	↑ **61.8**	67.4	57.5	**62.1**	↓ 64.1	↑ 61.6	↑ **62.9**
fr, nl, sl	64.9	58.2	**61.4**	↓ 61.5	↑ 61.0	↓ 61.3	65.3	57.9	61.4	↓ 63.1	↑ **62.7**	↑ **62.9**
fr, en, nl	68.0	50.7	58.1	↓ 65.4	↓ 46.9	↓ 54.6	70.2	52.1	59.8	↓ 63.8	↑ 56.5	↑ 60.0
en, fr, nl, sl	58.1	**61.6**	59.8	↑ 60.3	↑ **62.8**	↑ 61.6	59.5	**62.5**	61.0	↑ 64.2	↓ 59.5	↑ 61.7
cross-ling. avg.	63.1	54.4	58.0	↓ 62.3	↓ 53.3	↓ 56.9	64.6	57.3	60.4	↓ 64.5	↓ 55.3	↓ 59.4
multi-ling. avg.	64.3	55.2	59.2	↓ 63.0	↑ 55.9	↓ 58.9	66.6	55.7	60.5	↓ 63.6	↑ 58.3	↑ 60.6
Dutch test set.												
nl	70.3	62.2	66.0	↑ 71.2	↑ 64.1	↑ 67.5	73.3	61.5	66.9	↑ **73.5**	↑ 62.6	↑ 67.6
en	69.2	61.1	64.9	↑ 71.0	61.1	↑ 65.7	73.0	63.0	67.6	↓ 69.4	↑ 68.4	↑ 68.9
fr	72.1	51.0	59.8	↓ 70.4	↑ 55.6	↑ 62.2	73.6	55.5	63.3	↓ 70.4	↑ 62.4	↑ 66.2
en, sl	59.5	76.6	67.0	↑ 61.6	↑ 78.3	↑ 68.9	61.1	73.6	66.7	↑ 61.6	↑ **75.7**	↑ 68.4
fr, sl	62.5	74.7	68.1	↓ 58.7	↑ **79.3**	↓ 67.5	61.6	71.2	66.1	↓ 59.4	↑ 75.1	↑ 66.3
en, fr	72.5	61.7	66.7	↓ 70.8	↓ 60.1	↓ 65.0	73.1	63.5	68.0	↓ 72.5	↓ 61.2	↓ 66.4
en, fr, sl	59.6	**77.0**	67.2	↑ 61.1	↑ 78.2	↑ 68.6	66.6	69.6	68.1	↓ 66.4	↑ 74.9	↑ 70.4
nl, en	69.3	60.2	64.4	↓ 68.6	↑ 62.7	↑ 65.5	74.4	61.7	67.4	↓ 70.7	↑ 66.3	↑ 68.4
nl, fr	**75.7**	56.7	64.8	↓ **73.2**	↑ 58.1	64.8	**76.7**	59.6	67.1	↓ 73.0	↑ 60.6	↓ 66.2
nl, sl	65.8	72.7	**69.1**	↓ 65.0	↑ 77.0	↑ 70.5	69.9	69.7	**69.8**	↓ 68.6	↑ 72.5	↑ 70.5
nl, en, sl	64.7	73.0	68.6	↓ 60.0	↑ 80.6	↑ 68.8	68.7	70.3	69.5	↓ 67.6	↑ 74.2	↑ 70.8
nl, fr, sl	69.2	69.0	**69.1**	↓ 65.2	↑ 76.5	↑ 70.4	69.4	69.4	69.4	↓ 65.4	↑ 74.4	↑ 69.6
nl, en, fr	69.9	64.3	67.0	↑ 72.1	↓ 55.5	↓ 62.7	73.7	62.9	67.9	↓ 71.1	↑ 64.9	↓ 67.8
en, fr, nl, sl	62.7	75.5	68.5	↑ 64.5	↑ 78.1	↑ **70.6**	63.6	**73.7**	68.3	↑ 69.2	↓ 73.2	↑ **71.1**
cross-ling. avg.	65.9	67.0	65.6	↓ 65.6	↑ 68.8	↑ 66.3	68.2	66.1	66.6	↓ 66.6	↑ 69.6	↑ 67.8
multi-ling. avg.	68.2	67.3	67.4	↓ 66.9	↑ 69.8	↑ 67.6	70.9	66.8	68.5	↓ 69.4	↑ 69.4	↑ 69.2

4 Results

In this Section, we determine the predictive power of monolingual, cross-lingual, and multilingual learning in ACTER and RSDO5 test sets as well as compare the results from our proposed approaches to the SOTA from the related work.

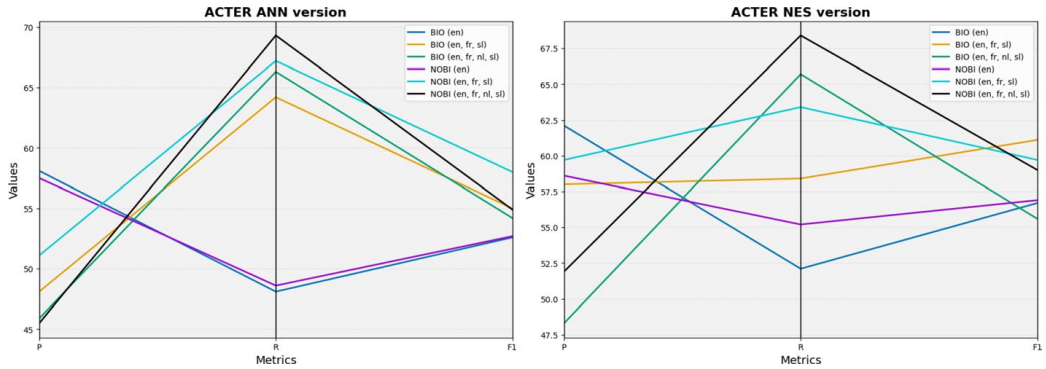

Fig. 2 Parallel Coordinates Plot in performance of XLMR classifier for the English test set

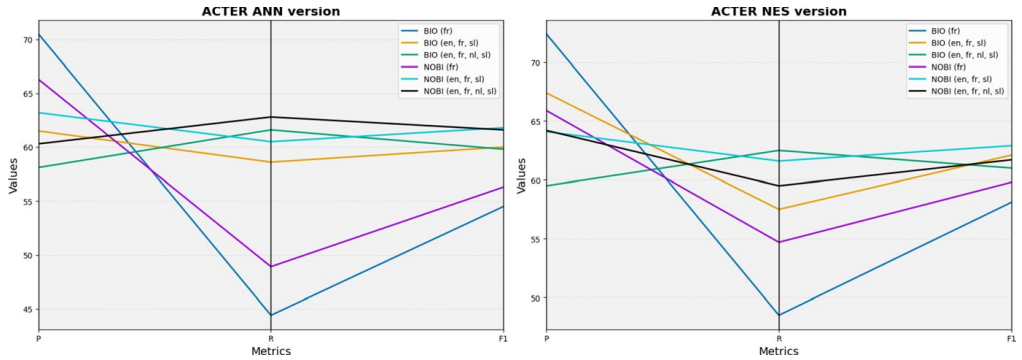

Fig. 3 Parallel Coordinates Plot in performance of XLMR classifier for the French test set

4.1 Results on the ACTER test set

The performance of the XLMR classifier regarding P, R, and F1 on the ACTER test set using BIO and NOBI annotation regimes are presented in Table 1. The comparison between BIO and NOBI is indicated with arrows, where ↑ is used to show better performance of NOBI in the same setting, while ↓ denotes lower performance. No matter which annotation scheme, the results indicate that the cross-lingual and multilingual models in both versions of test data, where one excludes the named entities of the test data (ANN) and the other includes them (NES), tend to surpass the performance of the monolingual ones according to all evaluation metrics, except for the Precision obtained by the French monolingual model on the French test set when the BIO scheme is used and Dutch monolingual model on the Dutch test set when NOBI scheme is used.

Multilingual models tend to outperform cross-lingual ones in F1. However, multilingual models have a tendency to lose their competency in Precision toward monolingual and cross-lingual ones. By adding the Slovenian corpus with four different domains into the training set, the multilingual model demonstrates a significant improvement in Recall across all test languages compared with the monolingual setting. It also outperforms other models in the F1 when we evaluate it in all three test sets in both annotation schemes. However, this improvement is at the cost of Precision.

When it comes to the comparison of the two annotation regimes, using the NOBI annotations in many cases improves the Recall of the model. This is especially visible in

Table 2 F1 comparison between our XLMR classifier in multilingual settings and related work in ACTER corpora

Methods	English		French		Dutch	
	ANN	NES	ANN	NES	ANN	NES
Winning teams (Hazem et al. 2020)	45.0	46.7	45.9	48.2	18.6	18.7
HAMLET (Rigouts et al. 2021)	54.2	55.4	60.2	60.8	66.1	66.0
Sequence Classifier (Lang et al. 2021)	x	46.0	x	48.1	x	58.0
NMT (Lang et al. 2021)	x	55.3	x	57.6	x	59.6
Token classifier (Lang et al. 2021)	x	58.3	x	57.6	x	69.8
NMF-based approaches (Nugumanova et al. 2022)	33.5	33.7	30.9	30.7	30.1	30.3
BIO classifier	54.9	**59.7**	61.4	62.1	69.1	69.8
NOBI classifier	**58.0**	59.3	**61.8**	**62.9**	**70.6**	**71.1**

Bold indicates the best result for each test set

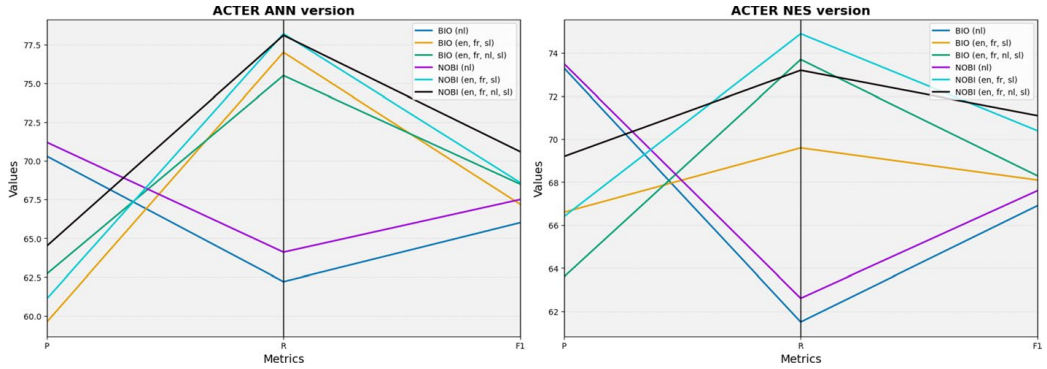

Fig. 4 Parallel Coordinates Plot in performance of XLMR classifier for the Dutch test set

the monolingual and multilingual settings (see Figs. 2, 3, and 4) in which the models are trained in multiple languages including the language of the test sets for all scenarios, and cross-lingual settings in which the models are trained on just one language and applied to the others except for French test set. A substantial increase in Recall also tends to lead to the improvement of the overall F1.

The best models from our combinations include: (1) For the English and French test sets, the best results were obtained with English, French, and Slovenian training data; and (2) For the Dutch test set, the best results were gained with the multilingual classifiers of all four languages. Thus, we compare the multilingual XLMR classifier fine-tuned on the pre-defined test language and multiple languages (trained in at least three languages including Slovenian and the test set's language) using the ACTER dataset in both annotation regimes. This showcases the power of a multilingual pretrained language model with multilingual settings - using (1) English, French, and Slovenian; and (2) all four languages as the training set - in capturing and understanding diverse linguistic nuances in comparison with a monolingual one. Additionally, the NOBI regime outperforms BIO ones for most of the testing scenarios.

Besides, we also compare the proposed results with the benchmarks as in Table 2 to highlight our hypothesis. For comparison, we include the solutions from the winning

Table 3 The evaluation in RSDO5 corpus given each domain as a test set in monolingual setting. Bold indicates the best result for each test set. The comparison between BIO and NOBI as well as the best model in F1-score are set in the same mechanism with Table 1

Valid set	Test set	BIO P	BIO R	BIO F1	NOBI P	NOBI R	NOBI F1
vet	ling	**69.6**	64.1	66.7	↓ 65.4	↑65.4	↓ 65.4
bim	ling	69.5	**73.7**	**71.5**	↓ **66.9**	↓ 69.5	↓ 68.2
kem	ling	66.2	72.4	69.2	↓ 64.9	↓ **72.3**	↓ **68.4**
ling	vet	71.1	66.7	68.8	↓ 66.6	↑68.5	↓ 67.5
kem	vet	**72.7**	65.6	**68.9**	↓ 66.9	↑ **69.7**	↓ **68.3**
bim	vet	69.3	**68.1**	68.7	↓ **67.6**	↓ 62.5	↓ 65.0
ling	kem	68.7	55.1	61.2	↓ 63.8	↑ **61.4**	↑ 62.6
bim	kem	**70.2**	**60.3**	**64.8**	↓ 66.1	↑ **61.4**	↓ 63.7
vet	kem	**70.2**	59.2	64.3	↓ **68.3**	↑ 60.6	↓ **64.2**
vet	bim	**63.5**	**66.8**	**65.1**	↓ **61.4**	↓ 61.3	↓ **61.3**
ling	bim	62.3	65.2	63.7	↓ 57.2	↓ 60.1	↓ 58.6
kem	bim	62.4	64.0	63.2	↓ 61.0	↓ **61.7**	↓ **61.3**
Avg.		68.0	65.1	66.3	↓ 64.7	↓ 64.5	↓ 64.5

Table 4 The evaluation in RSDO5 corpus given each domain as a test set in the multilingual setting. In this setting, in addition to Slovenian training data, the data from ACTER in en, fr, and nl is used, and ANN and NES training sets are compared

Valid. set	Test set	ANN BIO P	ANN BIO R	ANN BIO F1	ANN NOBI P	ANN NOBI R	ANN NOBI F1	NES BIO P	NES BIO R	NES BIO F1	NES NOBI P	NES NOBI R	NES NOBI F1
vet	ling	67.7	69.6	68.6	↓ **67.5**	↓ 62.7	↓ 65.0	67.2	**69.9**	**68.5**	↓ 64.2	↓ **67.3**	↓ **65.7**
bim	ling	**69.8**	66.2	67.9	↓ 64.6	↓ 68.1	↑ **66.3**	67.8	68.5	68.2	↓ **64.9**	↓ 64.8	↓ 64.8
kem	ling	66.5	**71.4**	**68.8**	↓ 59.6	↓ **71.0**	↓ 64.8	**67.9**	69.0	**68.5**	↓ 59.9	↓ 65.1	↓ 62.4
ling	vet	**71.0**	65.3	68.0	↓ 62.4	↑ **70.9**	↓ 66.4	69.2	67.4	68.3	↓ 61.8	↑ 70.8	↓ 66.0
kem	vet	69.8	**68.8**	**69.3**	↓ 68.0	↓ 68.5	↓ **68.2**	**70.5**	**67.8**	**69.1**	↓ **64.6**	↑ 70.6	↓ **67.5**
bim	vet	69.8	68.4	69.1	↓ **68.7**	↓ 67.1	↓ 67.9	69.3	64.7	66.9	↓ 63.0	↑ **72.8**	↓ **67.5**
ling	kem	68.3	59.3	63.5	↓ 66.0	↓ 52.9	↓ 58.7	67.5	54.6	60.4	↓ 62.8	↑ **60.8**	↑ 61.8
bim	kem	69.6	**61.2**	**65.1**	↓ **66.6**	↓ 55.5	↓ 60.5	**69.3**	52.7	59.9	↓ **65.5**	↑ **60.8**	↑ **63.1**
vet	kem	**69.9**	58.4	63.6	↓ 65.9	↓ **57.7**	↓ **61.5**	67.9	**59.2**	**63.3**	↓ 62.8	↑ **60.8**	↓ 61.8
vet	bim	61.2	**64.9**	**63.0**	↑ **62.9**	↓ 62.6	↓ 62.7	60.9	66.7	63.7	↓ 59.1	↓ 64.0	↓ 61.5
ling	bim	60.5	63.8	62.1	↓ 56.2	↓ 58.2	↓ 57.2	62.6	62.3	62.4	↓ 57.0	↑ 62.9	↓ 59.8
kem	bim	**65.7**	59.2	62.3	↓ 59.5	↑ **66.7**	↑ **62.9**	61.8	**67.1**	**64.3**	↓ 61.0	**67.1**	↓ **63.9**
Avg.		67.5	64.7	65.9	↓ 64.0	↓ 63.5	↓ 63.5	66.8	64.2	65.3	↓ 62.2	↑ 65.6	↓ 63.8

teams in the competition (TALN-LS2N (Hazem et al., 2020) won on the English and French test set, while NLPLab UQAM (Le & Sadat, 2021) won on the Dutch test set) and other methods (Rigouts et al., 2021; Lang et al., 2021) described in Sect. 2. Note that all the approaches from the related work are (1) cross-domain and (2) use the Heart failure domain as the test set, which shares the same mechanism with our approaches' validation.

Our proposed classifiers, trained using either BIO or NOBI annotation regimes, outperform previously described benchmark approaches, showcasing significant performance gains as measured by the F1. When comparing classifiers using BIO and NOBI annotation schemes, those utilizing BIO regimes demonstrate superior F1 on the English NES gold standard, which includes named entities. However, classifiers employing NOBI regimes exhibit noteworthy performance, surpassing all existing state-of-the-art (SOTA) models, including our BIO classifiers, across the languages present in both ANN and NES versions, with the exception of the aforementioned English NES corpus.

Machine Learning (2024) 113:4285–4314 4297

Table 5 Comparison between our performance and SOTA in RSDO5 dataset

Methods	Linguistics			Veterinary			Chemistry			Biomechanics		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SOTA Ljubešić et al. (2019)	52.2	25.4	34.1	66.9	19.3	29.9	47.8	31.4	37.8	53.8	24.8	33.9
Mono BIO	69.5	73.7	71.5	72.7	65.6	68.9	70.1	60.3	64.8	63.5	66.8	65.1
Multi BIO	66.5	71.5	68.8	69.8	**68.8**	**69.3**	69.6	**61.2**	**65.1**	61.8	**67.1**	64.3
Mono NOBI	64.9	72.3	68.4	66.9	69.7	68.3	68.3	60.6	64.2	61.4	61.3	61.3
Multi NOBI	64.6	68.1	66.3	68.0	68.5	68.2	65.5	60.8	63.1	61.0	67.1	63.9

Furthermore, we conduct a multilingual evaluation to examine the impact of adding additional languages to the training set. In contrast to the findings of Lang et al. (2021), we observe that incorporating other languages generally leads to only marginal improvements in model performance.

4.2 Evaluation on the RSDO5 test set

We also apply monolingual and multilingual cross-domain approaches to the Slovenian RSDO5 dataset. The results grouped by the test domain using BIO and NOBI annotation regimes are presented in Tables 3 and 4, respectively. For each annotation regime, we evaluate monolingual and multilingual settings where ANN and NES versions are added to the training set of the RSDO5 corpus.

The monolingual approach, where we use two domains from the RSDO5 corpus for training, validate on the third domain, and test on the last domain, proves to have relatively consistent performance across all the combinations in both annotation regimes. For both regimes, we achieve a Precision of more than 61%, Recall of no less than 55%, and F1 above 57%. Furthermore, they perform slightly better in the Linguistics and Veterinary domains than in Biomechanics and Chemistry. The difference in the number of terms and length of terms per domain pointed out in Sect. 3.1 might be one of the factors that contribute to this behavior. Moreover, a significant performance boost can be observed for the Veterinary domain when the model is trained in the Biomechanics and Linguistics domains and for the Linguistics domain if the Veterinary domain is included in the training set for the model in both annotation regimes. Between these two settings, the classifier with BIO regime gained a performance of up to 68.9% in the F1 for the Linguistics test set, which surpasses other domains in the same regimes as well as outperforms all the cases in the monolingual classifier of the NOBI regime.

We also explore the performance of multilingual approaches on the RSDO5 test sets. We train the model using the ANN and NES labels from all domains of the ACTER dataset and on two domains from the RSDO5 dataset, validate on the third RSDO5 domain, and test on the last domain. Table 3 and 4 present the comparative performance of the multilingual and the monolingual approaches. However, from the results, there exists a discrepancy in the performance-boosting efficiency among the different combinations of training, validation, and test sets. This raises a hypothesis of the domain sensitivity in transfer learning for ATE tasks. Thus, a careful choice of the domains in the training set is undoubtedly necessary for boosting the classifier's performance.

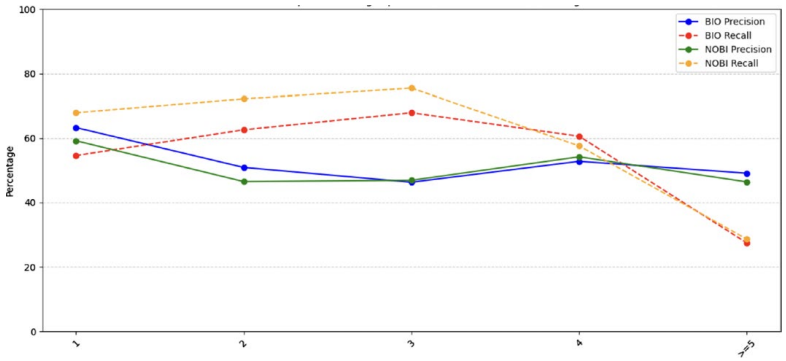

Fig. 5 Performance in P and R per term length per domain in English ACTER test set

Besides, we compare two different annotation regimes by evaluating the performance of classifiers using different training, validation, and testing combinations for each regime. Despite the consistency in the predictive power of monolingual and multilingual settings, the classifiers with NOBI annotation presented a worse performance in the Slovenian RSDO5 corpus compared to the BIO regime. This is due to the fact that the proportion of nested terms in RSDO5 is too small for the classifier to learn nested terms properly, which are visualized in the proportion of unique nested terms and terms nested in other nested terms from Figs. 16 to 18.

In Table 5, we present the results from the related work for the RSDO5 dataset compared to our proposed monolingual and multilingual approaches. The result from Ljubešić et al. (2019)'s method, which has been re-implemented using the same RSDO5 corpus as our studies, is taken from Tran et al. (2022b). In general, our approach outperforms Ljubešić et al. (2019)'s one by a large margin on all domains and according to all evaluation metrics, especially when it comes to Recall. We achieve results roughly twice as high as Ljubešić et al. (2019)'s approach in F1-score for all test domains regarding both monolingual and multilingual learning. One should note that the method (Ljubešić et al., 2019) was primarily meant for extracting terms from Ph.D. theses, i.e., documents significantly longer than those available in our training data, which explains the low Recall of that approach. However, this result clearly identifies a significant strength of the sequence-labeling approach - it does not rely on the frequency of term occurrences, which makes the approach more robust as shown in this comparison. In our case, we show that the multilingual experiments do in several cases improve our monolingual results (Tran et al., 2022b), but not systematically.

5 Error analysis

In order to determine whether the term length affects the models' performance, we calculate Precision and Recall for terms of length k = {1,2,3,4, ≥5} when predicted by our classifiers on the test set. The number of predicted candidate terms (Preds), number of ground truth terms (GTs), number of correct predictions (TPs), Precision (P), and Recall (R) regarding different term lengths k and test domains in ACTER and

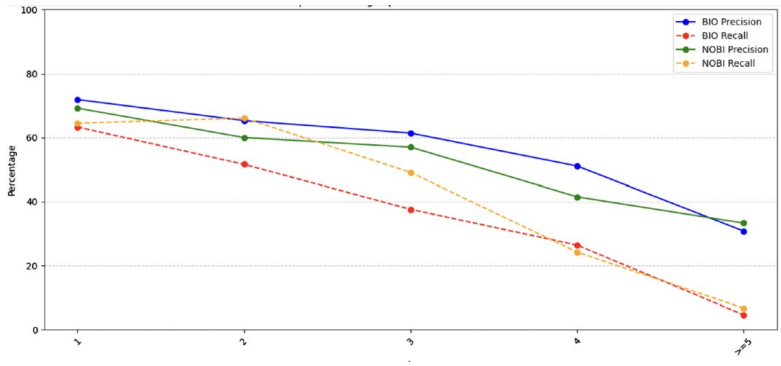

Fig. 6 Performance in P and R per term length per domain in French ACTER test set

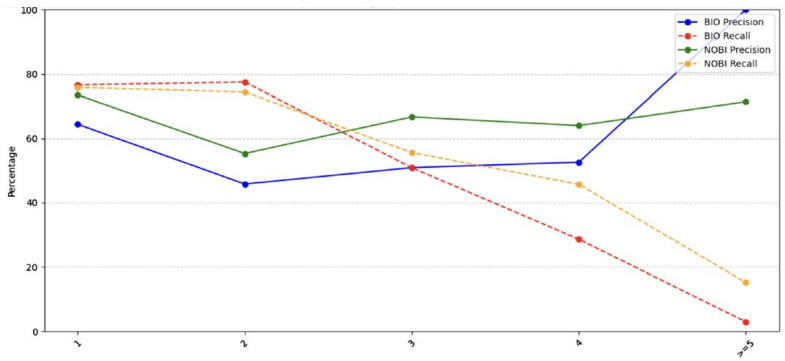

Fig. 7 Performance in P and R per term length per domain in Dutch ACTER test set

RSDO5 corpora are presented in Table 9 and 10 (in Appendix) and Precision (P) and Recall (R) of each scenario are visualized below.

5.1 The ACTER dataset

The results for ACTER's dataset (Table 9) were obtained by employing the best performing model for a specific language in terms of F1 on the Heart failure test set for the most cases (which is the combination of English, French, and Slovenian as the training set).

As demonstrated in Fig. 5, 6, and 7, when using the BIO scheme, the best model proved to be good at predicting terms containing up to four words for English and Dutch and up to three words for French texts in ACTER corpora. A strong correspondence between the F1 and the number of predicted candidate terms has been found where the number of predicted candidate terms likely corresponds to the situation in the training data (see Table 9 in Appendix).

The best models trained using the NOBI annotation scheme demonstrated the same behavior as the one trained using the BIO annotation regime. They performed well at predicting terms containing up to four words for English and Dutch and up to three words for French texts in ACTER corpora. While our expectation was that the NOBI annotation scheme should benefit the model's ability to predict short one-word nested terms, the classifiers trained using NOBI annotations show better performance than those using the BIO regime on multi-word terms as well, as long as nested terms take a proper proportion as

Table 6 A comparison of the performance between the BIO and NOBI schemes on the entire dataset, single-word terms (SWU), and multi-word terms (MWU)

	BIO			NOBI		
	P	R	F1	P	R	F1
All terms	58.1	48.1	52.6	57.5	48.6	52.7
SWU	65.0	45.9	53.8	61.6	51.5	56.1
MWU	53.8	50.0	51.8	54.2	46.3	49.9

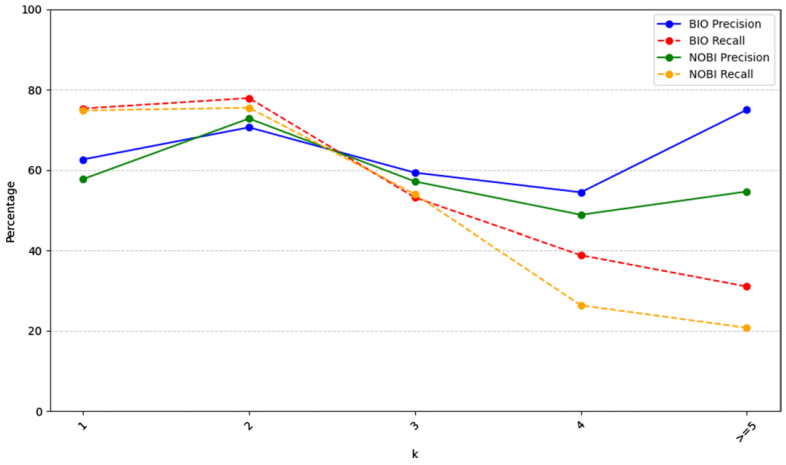

Fig. 8 Performance in P and R per term length per domain in RSDO Linguistics test set

in ACTER corpora. The Recall therefore generally improves for terms of all lengths, even for terms containing 5 words or more. There seems to be some signal in the occurrence of nested terms inside multi-word terms, which leads the model to better identify longer terms as well. Our current hypothesis is that this effect is a combination of (1) the improvement of single-word term identification by having a larger training set available (both nested and independent single-word terms) and (2) nested terms being some sort of anchor exploited by the model to easier identify multi-word terms around that nested terms. Further experiments and analyses should be conducted to fully understand this phenomenon.

Furthermore, a trend that is noticeable across the majority of scenarios is that the NOBI regime reduces the Precision compared to the BIO regime. This seems to be related to the number of terms predicted where we can observe that Precision often drops where the number of predicted terms is higher, i.e., the BIO regime on the English dataset predicts 1,009 single-word terms with a Precision of 63.3 % and the NOBI regime predicts 1,341 terms with a Precision of 59.2%. In a similar but reversed trend, the Dutch NOBI regime produced 1,738 terms with a Precision of 73.5% whereas the BIO regime produced 2005 terms with a Precision of 64.4% (see Table 9 for the statistics).

We performed an additional detailed comparison of the BIO and NOBI monolingual results on the English dataset (i.e., the results from the first line in Table 1) in Table 6. The NOBI scheme produces a marginal improvement in terms of F1 and Recall but has slightly lower Precision. Overall, the algorithm predicted 1,956 candidates when using the BIO scheme and 1,996 when using the NOBI scheme. Out of these, the BIO scheme resulted in 751 single-word terms (SWU) and 1205 multi-word terms (MWU), while the NOBI scheme produced 889 single-word terms and 1,107 multi-word terms. Looking at the performance in Table 6, NOBI results in a better Recall of single-word terms (51.5 vs. 45.9),

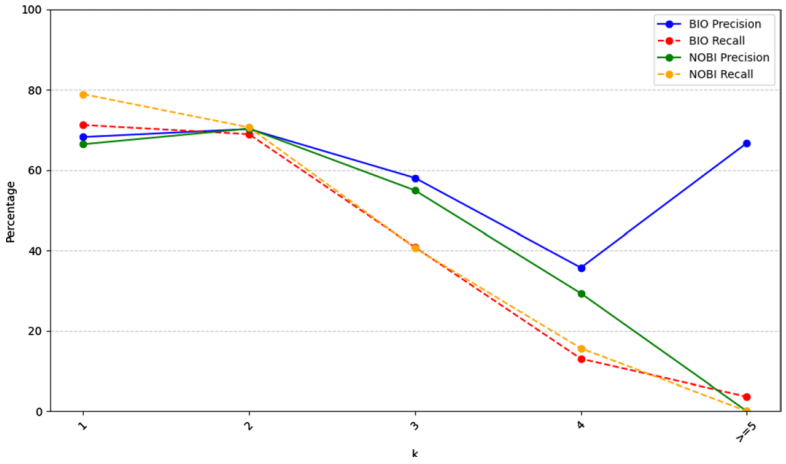

Fig. 9 Performance in P and R per term length per domain in RSDO Veterinary test set

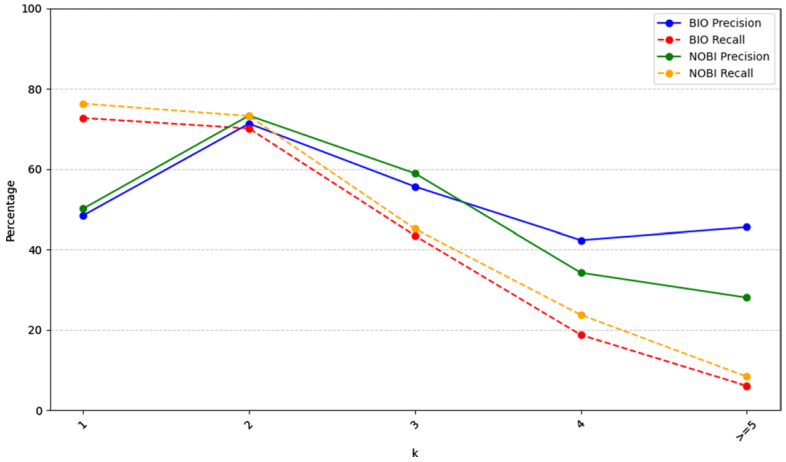

Fig. 10 Performance in P and R per term length per domain in RSDO Biomechanics test set

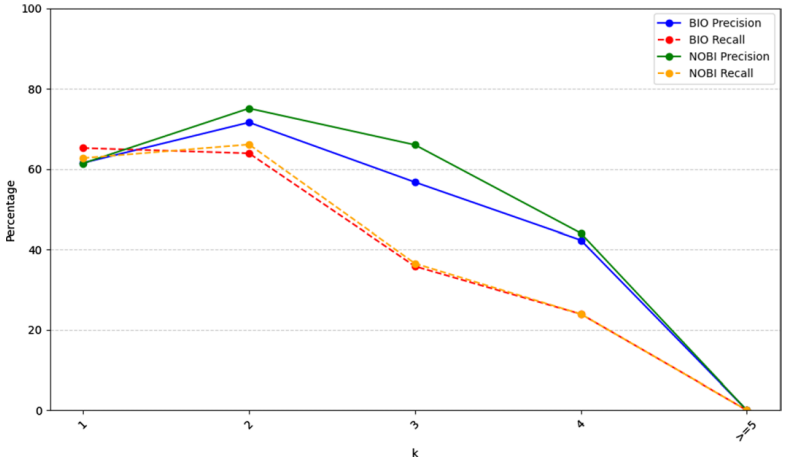

Fig. 11 Performance in P and R per term length per domain in RSDO Chemistry test set

which leads to an overall improvement of the F1 (52.7 vs. 52.6). It does not improve the Precision of SWU terms, but does, perhaps surprisingly, deliver higher Precision on MWU terms, which could be due to the fact that the NOBI regime prefers single-word terms (due to their higher proportion in the training set) which results in a smaller number of higher quality MWU terms being predicted.

5.2 The RSDO5 dataset

The results for the RSDO5 dataset (Table 10 in Appendix and from Figs. 8, 9, 10 and 11) were obtained by employing the best-performing model in the F1 for each specific test domain for both annotation regimes, which are (1) training on Veterinary and Chemistry, validation on Biomechanics, and testing on Linguistics domain; (2) training on Linguistics and Biomechanics, validation on Chemistry, and testing on Veterinary domain; (3) training on Linguistics and Veterinary, validation on Biomechanics, and testing on Chemistry domain; (4) training on Linguistics and Chemistry, validation on Veterinary, and testing on Biomechanics.

These results are similar to ACTER corpora, showing that the models are good at predicting short terms containing up to three words for all four domains of the Slovenian corpus. The best model applied to the Linguistics test domain also shows relatively good performance when it comes to the prediction of longer terms, achieving 75.0% Precision and a decent 31.0% Recall for terms with at least five words. Despite the relatively high Precision for prediction of long terms in the Veterinary and Biomechanics test domains, the Recall is pretty low, most likely due to the small amount of longer terms in the dataset on which the models are trained. When predicting the Chemistry domain, there are no correct predictions of more than five-word terms.

The NOBI regime often results in a lower Precision compared to the BIO one. Similar to our findings on the ACTER dataset, this seems to be related to the number of terms being predicted. In general, the higher the number of predictions, the lower the Precision (if the number of predicted terms is high enough — this trend is less noticeable for longer terms of which there are few in the corpus). There are some exceptions, like the Chemistry domain, where the NOBI regime results in 909 predicted single-word terms with a Precision of 61.4% compared to 943 terms with a Precision of 61.5% for the BIO regime, and the Veterinary domain where the NOBI regime predicted 2,111 two-word terms (k=2) with a Precision of 70.3% while the BIO regime predicted 2062 terms with a Precision of 70.2%.

As mentioned above, as well as in previous work (Tran et al., 2022b) for the BIO regime, since the corpus contains nested terms, the very common mistake the both BIO and NOBI models make is to incorrectly predict a shorter term nested in the correct term of the gold standard. Vice versa, the model sometimes generates incorrect predictions containing the correct nested terms. However, the NOBI annotation proves to partially reduce the effect of these two mentioned error patterns and improves the general Recall in comparison to the benchmark BIO scheme.

6 Conclusion

In summary, we demonstrated the possibilities of cross- and multilingual learning compared to the monolingual setting in boosting the predictive performance of the cross-domain sequence-labeling term extraction via experiments conducted on multi-domain

corpora, namely the ACTER and RSDO5 datasets. In addition, we presented the positive impact of cross- and multilingual models on the ACTER corpora only, and by further adding the texts from the Slovenian RSDO5 corpus in the training set. Furthermore, we examined the cross-lingual effect of rich-resourced training language on less-resourced testing ones such as Slovenian. Last but not least, we proposed a new NOBI annotation regime, that boosted the predictive power of classifiers in comparison to the classical BIO mechanism, as shown in the ACTER corpus, in which the number of nested terms is significant enough. The improvements through the NOBI annotation regime are visible even in multi-word term identification, quite likely by improving single-word term extraction and exploiting single-word terms as anchors to correctly identify multi-word terms. The results demonstrated the potential of the new annotation scheme to enhance the nested term extraction and a promising impact of cross- and multilingual cross-domain learning when transferring from rich- to less-resourced languages.

In future work, we will test the potential of our proposed NOBI mechanism in similar sequence-labeling extraction tasks in other domains (e.g., Named Entity Recognition). In addition, we plan to investigate the integration of active learning into our current approach to improve the output of the automated method by dynamical adaptation after human feedback.

Appendix

Data analysis

Figure 12 presents the structures of two datasets that we used for our work, including ACTER corpora and RSDO5 corpus. Note that in ACTER datasets, two versions of the gold standard were proposed: (1) ANN version covering only terms; and (2) NES version including both terms and named entities.

Figure 13 illustrates an example of the key difference between the ACTER's ANN and NES versions of gold standards. Given the sentence "*...This study uses the Medicare Patient Safety Monitoring System...*", the gold standard of the ANN version consists of only the term "*Patient*" as the only term was annotated as the ground truth. On the other hand, the NES version's gold standard includes the Named Entity (NE) "*Medicare Patient Safety Monitoring System*" as both domain-specific terms and NEs were annotated in the ground truth.

Figure 14 summarizes the number of unique terms (e.g., the term counts excluding the duplication) for each domain in both ACTER and the Slovenian RSDO5 corpora. It contains statistics for both ANN (annotating only terms) and NES (annotating both terms and named entities as the ground truth terms) versions in the ACTER set. This supports the statements in Subsection 3.1 and 3.2.

Table 7 indicates the proportion of the nested terms with different word length k where $k = \{1, 2, 3, 4, \geq 5\}$ for each domain and language of both corpora, which also supports to the statements in Subsection 3.3. The last column on the right calculates the percentage of single-word nested terms in total nested terms in the first level. On average, the amount of single-word nested terms accounts for 78.06% above all the nested terms on the first levels in the corpora. That is why we do not consider either multi-word nested terms or terms nested in other nested terms - so-called nested terms on the second or higher levels and we label single-word nested terms on the first level.

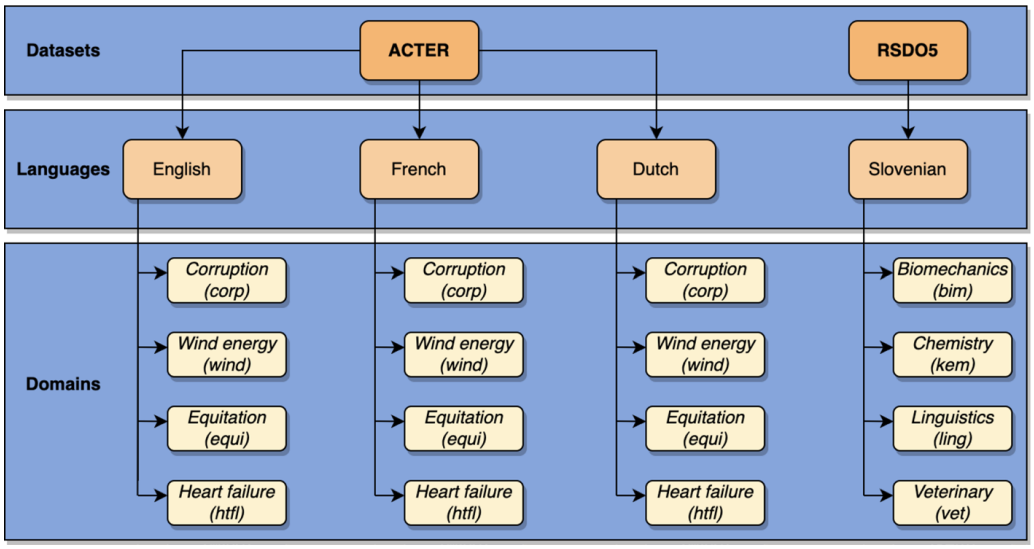

Fig. 12 The structure of RSDO5 and ACTER regarding languages and domains

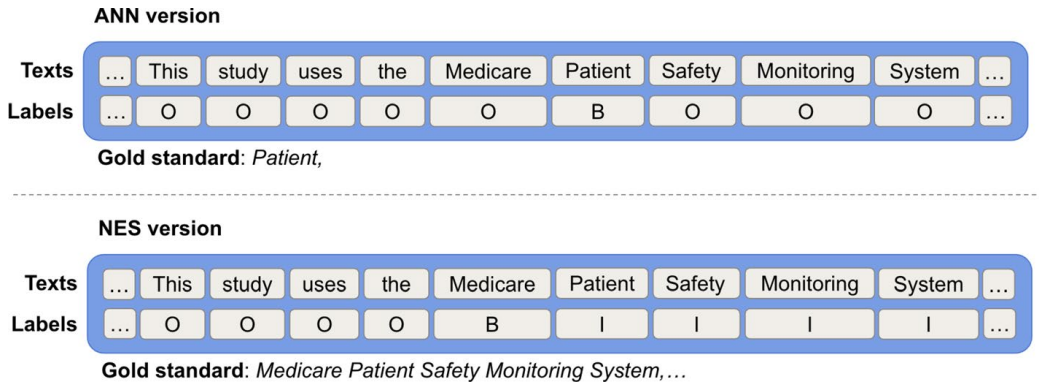

Fig. 13 An example of ACTER's ANN and NES versions were annotated in the BIO regime

Regarding ACTER corpora, Figs. 15 and 16 present the term density and the proportion of unique nested terms founded in texts extracted from the ACTER corpora for each domain and language, respectively. As can be seen from both figures, a notable disparity in data volume and term distribution is observed, particularly between the Heart failure domain and the other three domains, with the former containing a more significant number of unique terms. Further comprehensive information on the ACTER dataset can be found in the TermEval competition by Rigouts et al. (2020a).

Similarly, Figs. 17 and 18 present the term density and the unique term proportion in texts captured from the RSDO5 corpus for each domain, respectively. As can be seen, the documents from the Linguistics and Veterinary domains contain more terms than Biomechanics and Chemistry. Most terms are made of up to three words and only a few terms are longer than seven words. For example, an observation of the long multi-word term found in the corpus would be *"stojo po obračanju v nasprotni smeri urinega kazalca"* (stand after turning counterclockwise) in Biomechanics.

Machine Learning (2024) 113:4285–4314 4305

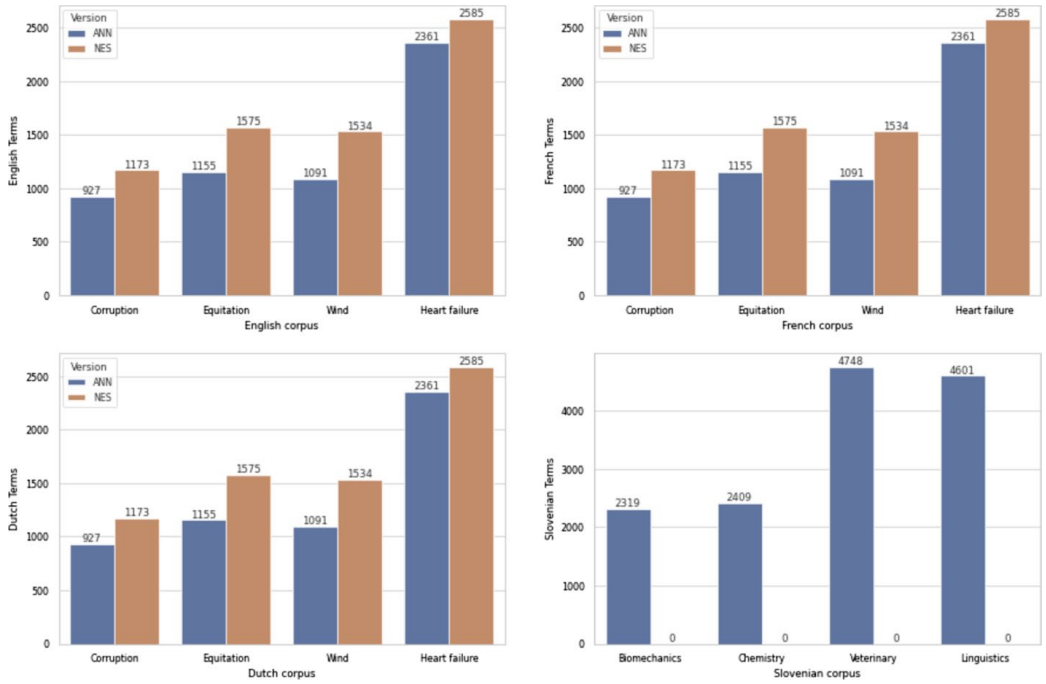

Fig. 14 The number of unique terms in ACTER and RSDO5 corpora

Table 7 The proportion of unique nested terms of different word lengths in each domain and language of ACTER and RSDO5 corpora

Languages	Domains	k = 1	k = 2	k = 3	k = 4	k ≥ 5	% (k = 1)
ACTER							
en	corp	246	89	11	1	1	70.69
	equi	469	87	5	1	0	77.90
	wind	282	171	36	4	0	83.51
	htfl	580	183	55	20	6	83.45
fr	corp	289	59	19	2	2	87.60
	equi	339	32	13	3	0	86.97
	wind	192	38	24	6	1	57.20
	htfl	620	99	30	8	9	73.56
nl	corp	309	46	12	2	1	84.90
	equi	414	44	12	6	0	68.72
	wind	253	36	4	4	1	80.94
	htfl	574	46	4	4	0	91.40
RSDO5							
sl	ling	737	177	8	0	0	79.93
	vet	835	199	13	5	1	79.30
	kem	388	126	7	2	1	74.05
	bim	349	111	17	16	14	68.84
Average							78.06

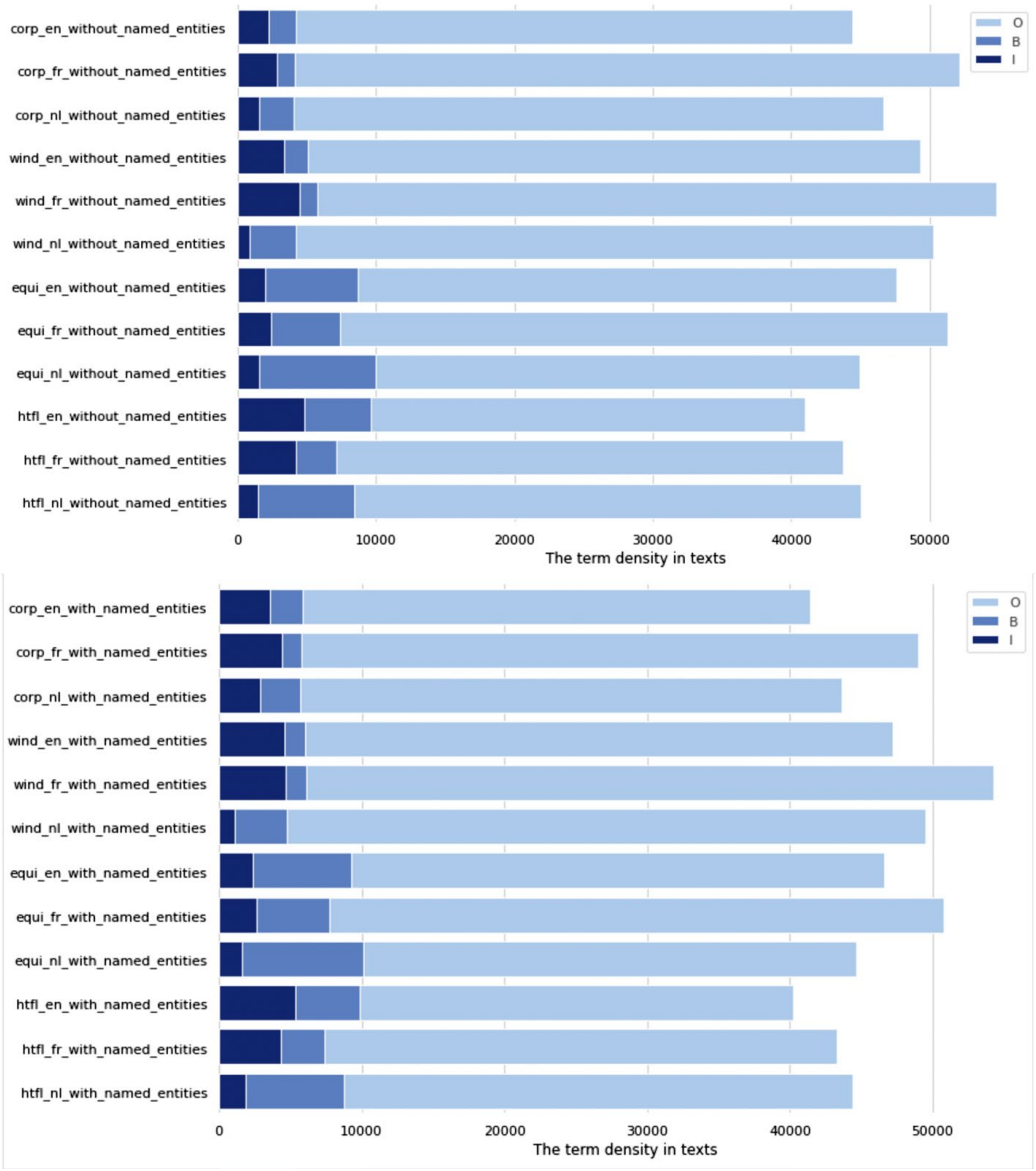

Fig. 15 The term density in BIO regime in the ACTER corpora

Annotation regimes

Besides the popular BIO regime, IOBES and BILOU are two different annotation schemes commonly used in Natural Language Processing (NLP) tasks. These schemes are used to represent and label entities within a sequence of words or tokens in a text. IOBES stands for tokens [**I**]nside an entity; [**O**]utside an entity (i.e., not part of any entity); [**B**]eginning token of an entity; [**E**]nd token of an entity; [**S**]ingle token that forms a whole entity by itself. Compared to the BIO scheme, the IOBES scheme is an extension of the BIO scheme with the "E" and "S" tags added to represent entities that end at a token or consist of a single token. Meanwhile, BILOU represents tokens [**B**]eginning token of an entity; [**I**]nside an entity; [**L**]ast token of an entity; [**O**]utside an entity; and [**U**]nit token that forms a whole entity by itself. Sharing the same "B", "I", and "O", the BILOU scheme is an extension

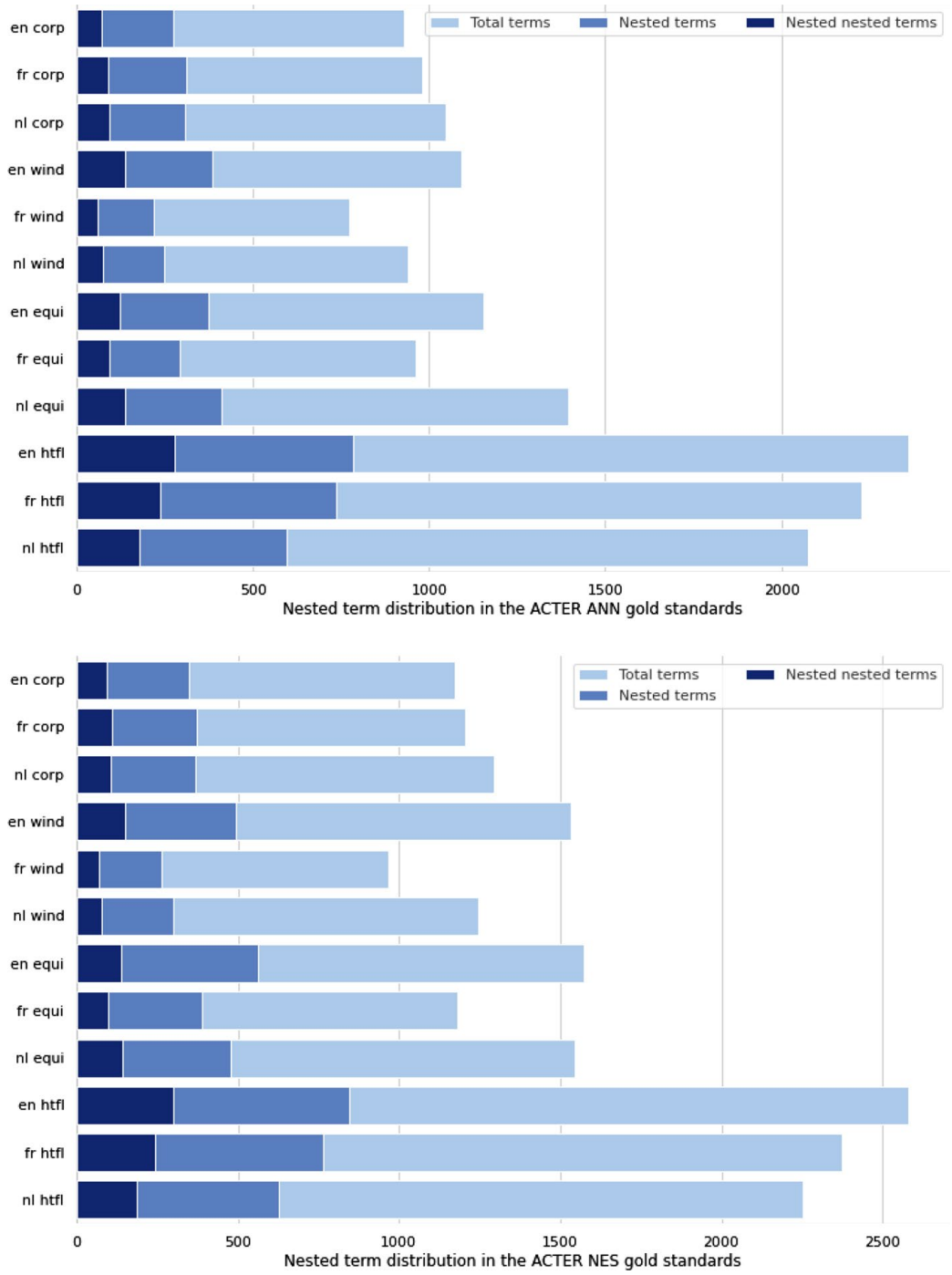

Fig. 16 The proportion of unique nested terms in the ACTER gold standards

of the IOBES scheme, but it offers a more compact representation of entities that consist of multiple tokens. We reported the performance of our XLMR classifier fine-tuning on ACTER English sets in BIO, NOBI, BIOES, and BILOU with ANN gold standard as demonstrated in Table 8.

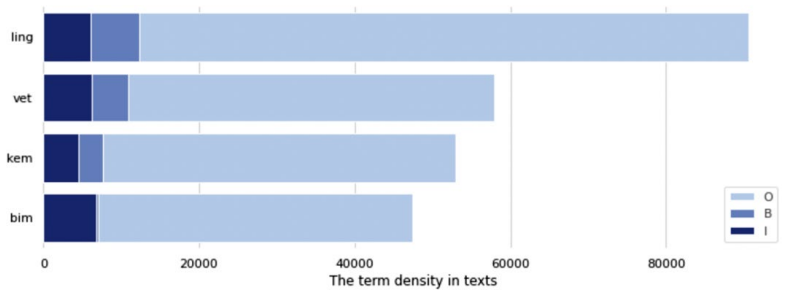

Fig. 17 The term density in BIO regime per domain of the RSDO5 corpus

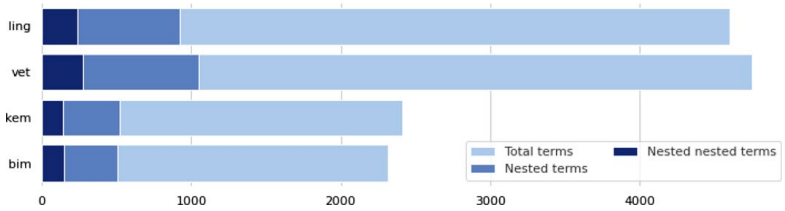

Fig. 18 The proportion of unique nested terms in the RSDO5 gold standards

The results demonstrate the superiority of our novel annotation regimes in comparison with other related schemes. In fact, both IOBES and BILOU are widely used to label entities and are often converted into simpler BIO formats during training or evaluation. These annotation schemes help models understand the boundaries and types of entities present in a text, enabling them to learn to recognize and extract them effectively. However, the standard IOBES and BILOU annotation schemes do not well support nested entities but were used as a foundation (similar to BIO) to improvise on the nested terms. Both IOBES and BILOU are designed to represent terms or entities in a flat manner, where each token in the text is associated with only one entity tag. In a nested entity scenario, we would have ones that are hierarchically structured, with one entity/term fully or partially contained within another entity/term. To represent nested entities, we propose more custom annotation schemes, namely NOBI, and a simple designed scheme to handle the single-word nested structures appropriately.

Monolingual vs. multilingual pre-trained models

We evaluated the performance using monolingual language models, including XLNet[5] (English), CamemBERT[6] (French), and DutchBERT[7] (Dutch) compared against a multilingual model, XLMR[8], which was pre-trained on over 100 different languages and fine-tuned for the downstream ATE task, as visualized from Figs. 19, 20, 21. The selection of the monolingual models is based on their superior performance in the empirical evaluation of various monolingual and multilingual Transformer-based models on monolingual sequence-labeling cross-domain term extraction (Tran et al., 2022c).

[5] xlnet-base-cased (https://huggingface.co/xlnet-base-cased).
[6] camembert-base (https://huggingface.co/camembert-base).
[7] GroNLP/bert-base-dutch-cased (https://huggingface.co/GroNLP/bert-base-dutch-cased).
[8] xlm-roberta-base (https://huggingface.co/xlm-roberta-base).

Table 8 Evaluation of XLMR classifier fine-tuning on ACTER English sets in BIO, NOBI, BIOES, and BILOU with NES gold standard

Models	P	R	F1
BIO	62.1	52.1	56.7
BIOES	62.6	51.9	56.7
BILOU	61.8	52.6	56.8
NOBI	58.6	55.2	56.9

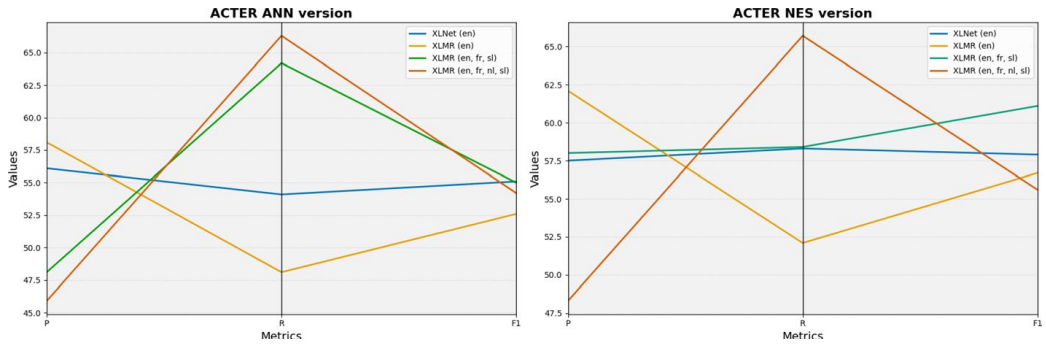

Fig. 19 Performance of monolingual pre-trained classifier finetuned on English test language vs. multilingual one finetuned on the test language and multiple languages in ACTER

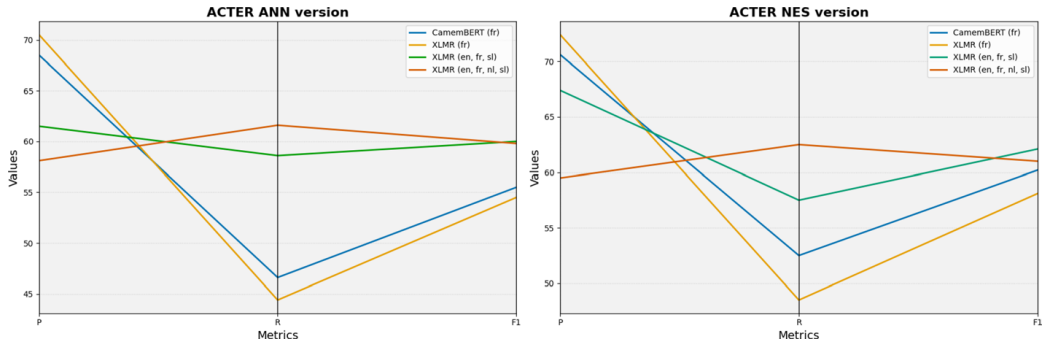

Fig. 20 Performance of monolingual pre-trained classifier finetuned on French test language vs. multilingual one finetuned on the test language and multiple languages in ACTER

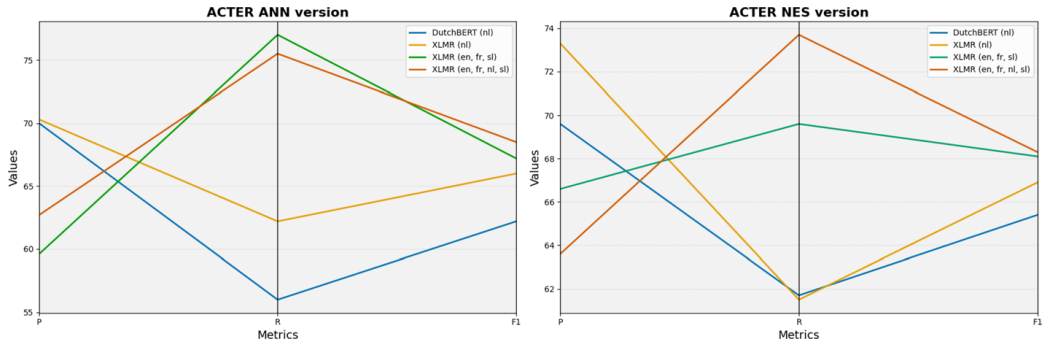

Fig. 21 Performance of monolingual pre-trained classifier finetuned on Dutch test language vs. multilingual one finetuned on the test language and multiple languages in ACTER

Table 9 Performance per term length per domain in ACTER dataset

k	English					French					Dutch				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
BIO regime															
1	1,009	1,170	639	63.3	54.6	1,153	1,309	829	71.9	63.3	2,005	1,687	1,292	64.4	76.6
2	985	801	501	50.9	62.6	490	620	320	65.3	51.6	661	391	303	45.8	77.5
3	553	377	256	46.3	67.9	163	266	100	61.4	37.6	108	108	55	50.9	50.9
4	163	142	86	52.8	60.6	47	91	24	51.1	26.4	19	35	10	52.6	28.6
≥5	53	95	26	49.1	27.4	13	88	4	30.8	4.6	1	33	1	100.0	3.0
NOBI regime															
1	1,341	1,170	794	59.2	67.9	1,219	1309	844	69.2	64.5	1,738	1,687	1,278	73.5	75.8
2	1,242	801	578	46.5	72.2	683	620	410	60.0	66.1	526	391	291	55.3	74.4
3	606	377	284	46.9	75.5	228	266	130	57.0	49.1	90	108	60	66.7	55.6
4	153	142	83	54.2	57.6	53	91	22	41.5	24.2	25	35	16	64.0	45.7
≥5	56	95	26	46.4	28.6	18	88	6	33.3	6.7	7	33	5	71.4	15.2

The results using the monolingual models exhibit slightly higher performance in the specific language they were pre-trained on. However, when applied in a cross-lingual context (e.g., fine-tuning XLNet on an English corpus and predicting on a French test set), the performance is significantly diminished when compared to the multilingual pre-trained model (e.g., XLMR). While the difference between the language-specific and multilingual models is small, the multilingual models, trained with XLMR on the datasets of multiple and all languages, for the most part, outperform the monolingual models by a small margin. As a result, in order to accommodate and support multiple languages simultaneously, we opt to utilize XLMR as the benchmark model for all four languages in ACTER and RSDO5 corpora to validate our hypothesis in this study.

Error analysis

We calculate Precision and Recall for terms of length k = {1,2,3,4, ≥5} when our classifiers predict on the test set. The number of predicted candidate terms (Preds), number of ground truth terms (GTs), number of correct predictions (TPs), Precision (P), and Recall (R) regarding different term lengths k and test domains in ACTER and RSDO5 corpora are presented in Table 9 and 10.

These Tables provide detailed support for the explanation of the classifier's behavior toward each dataset in terms of term length.

Table 10 Performance per term length per domain in RSDO corpus

NOBI regime										
k	Linguistics					Veterinary				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	2,078	1,728	1,300	62.6	75.3	2,159	2,067	1,472	68.2	71.2
2	2,631	2,404	1,858	70.6	77.9	2,062	2,103	1,448	70.2	68.9
3	322	360	7,191	59.3	53.1	314	446	182	58.0	40.8
4	57	80	31	54.4	38.8	28	77	10	35.7	13.0
≥5	12	29	79	75.0	31.0	3	55	2	66.7	3.6
k	Chemistry					Biomechanics				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	943	890	580	61.5	65.2	1,079	718	22	48.4	72.7
2	1,073	1,202	768	71.6	63.9	1,153	1,172	822	71.3	70.1
3	164	260	93	56.7	35.8	223	286	124	55.6	43.4
4	26	46	11	42.3	23.9	26	59	11	42.3	18.7
≥5	3	11	0	0.0	0.0	11	84	5	45.5	6.0
NOBI regime										
k	Linguistics					Veterinary				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	2241	1728	1293	57.7	74.8	2456	2067	1630	66.4	78.9
2	2491	2404	1814	72.8	75.5	2111	2103	1484	70.3	70.6
3	340	360	194	57.1	53.9	330	446	181	54.9	40.6
4	43	80	21	48.8	26.3	41	77	12	29.3	15.6
≥5	11	29	6	54.6	20.7	5	55	0	0.0	0.0
k	Chemistry					Biomechanics				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	909	890	558	61.4	62.7	1094	718	548	50.1	76.3
2	1058	1202	795	75.1	66.1	1171	1172	858	73.3	73.2
3	144	260	95	66.0	36.5	219	286	129	58.9	45.1
4	25	46	11	44.0	23.9	41	59	14	34.2	23.7
≥5	0	11	0	0	0.0	25	84	7	28.0	8.3

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Hanh Thi Hong Tran and Matej Martinc. The baseline method was proposed by Nikola Ljubesic. The first draft of the manuscript was written by Hanh Thi Hong Tran and all authors commented on previous versions of the manuscript. Andraz Repar proposed the analysis with respect to the difference between NOBI and BIO behavior in predicting the candidate terms. All authors read and approved the final manuscript.

Funding The work was partially supported by the Slovenian Research Agency (ARIS) via the core research programs Knowledge Technologies (P2-0103) and Language resources and technologies for Slovene (P6-0411), project Formant Combinatorics in Slovenian (J6-3131), as well as by the Ministry of Culture of

the Republic of Slovenia through the project Development of Slovene in Digital Environment (RSDO). The first author was partly funded by Region Nouvelle Aquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France. The work was also supported by the project Cross-lingual and cross-domain methods for Terminology Extraction and Alignment, a bilateral project funded by the program PROTEUS under the grant number BI-FR/23-24-PROTEUS006.

Data availability and materials The original datasets are collected from two sources: ACTER version 1.5 (https://github.com/AylaRT/ACTER) and RSDO version 1.1 (https://www.clarin.si/repository/xmlui/handle/11356/1470). The newly annotated dataset is publicly available at https://github.com/honghanhh/nobi_annotation_regime.git.

Code availability Our code is publicly available at https://github.com/honghanhh/ate_nobi.git.

Declarations

Conflict of interest Not applicable.

Ethical approval Not applicable.

Consent to participate All the authors consent to participate.

Consent for publication All the authors consent for publication.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (pp. 54–59).

Amjadian, E., Inkpen, D., Paribakht, T., & Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)* (pp. 2–11).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL*.

Daille, B., Gaussier, É., & Langé, J. M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Damerau, F. J. (1990). Evaluating computer-generated domain-oriented vocabularies. *Information Processing and Management, 26*(6), 791–801.

ElKishky, A., Song, Y., Wangx, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment, 8*(3), 305–316.

Erjavec, T., Fišer, D., & Ljubešić, N. (2021). The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation, 55*(2), 551–583.

Fišer, D., Suchomel, V., & Jakubícek, M. (2016). Terminology extraction for academic slovene using sketch engine. In *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016* (pp. 135–141).

Frantzi, K.T., Ananiadou, S., & Tsujii, J. (1998). The c-value/nc-value method of automatic recognition for multi-word terms. In *International conference on theory and practice of digital libraries* (pp. 585–604). Springer.

Gao, Y., & Yuan, Y. (2019). Feature-less End-to-end Nested Term extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 607–616). Springer.

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2020). TermEval 2020: TALN-LS2N System for Automatic Term Extraction. In *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 95–100).

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2022). Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 648–662).

Jemec Tomazin, M., Trojar, M., Atelšek, S., Fajfar, T., Erjavec, T., & Žagar Karer, M. (2021). Corpus of term-annotated texts RSDO5 1.1. URL http://hdl.handle.net/11356/1470. Slovenian language resource repository CLARIN.SI

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1*(1), 9–27.

Kessler, R., Béchet, N., & Berio, G. (2019). Extraction of terminology in the field of construction. In *2019 First International Conference on Digital Data Processing (DDP)* (pp. 22–26). IEEE.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., & Stüker, S. (2018). Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In *INTERSPEECH* (pp. 2072–2076).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 260–270).

Lang, C., Wachowiak, L., Heinisch, B., & Gromann, D. (2021). Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 3607–3620).

Le, N. T., & Sadat, F. (2021). Multilingual automatic term extraction in low-resource domains. In *The International FLAIRS Conference Proceedings*, vol. 34.

Le Serrec, A., L'Homme, M. C., Drouin, P., & Kraif, O. (2010). Automating the compilation of specialized dictionaries: Use and analysis of term extraction and lexical alignment. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication, 16*(1), 77–106.

Lester, B. (2020). iobes: A library for span-level processing. arXiv preprint arXiv:2010.04373.

Lingpeng, Y., Donghong, J., Guodong, Z., & Yu, N. (2005). Improving retrieval effectiveness by using key terms in top retrieved documents. In *European Conference on Information Retrieval* (pp. 169–184). Springer.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics, 8*, 726–742.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). Kas-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. In *International Conference on Text, Speech, and Dialogue* (pp. 115–126). Springer.

Marciniak, M., & Mykowiecka, A. (2015). Nested term recognition driven by word connection strength. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication, 21*(2), 180–204.

Martinc, M., Škrlj, B., & Pollak, S. (2021). Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering* (pp. 1–40). https://doi.org/10.1017/S1351324921000127

Nugumanova, A., Akhmed-Zaki, D., Mansurova, M., Baiburin, Y., & Maulit, A. (2022). NMF-based approach to automatic term extraction. *Expert Systems with Applications, 199*, 117179.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostaja, T., Vintar, Š., & Fišer, D. (2019). Extracting data from comparable corpora. In *Using Comparable Corpora for Under-Resourced Areas of Machine Translation* (pp. 89–139). Springer.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. *Natural language processing using very large corpora* (pp. 157–176).

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)* (pp. 147–155).

Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., & Pollak, S. (2019). TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment. *International Journal of Theoretical and Applied Issues in Specialized Communication, 25*(1), 93–120.

Rigouts Terryn, A., Hoste, V., Drouin, P., & Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)* (pp. 85–94). European Language Resources Association (ELRA).

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2020). In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation, 54*(2), 385–418.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2021). HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology*.

Tran, H. T. H., Doucet, A., Sidere, N., Moreno, J. G., & Pollak, S. (2021). Named entity recognition architecture combining contextual and global. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021*, Virtual Event, December 1–3, 2021, Proceedings, p. 264. Springer Nature.

Tran, H. T. H., Martinc, M., Doucet, A., & Pollak, S. (2022). Can cross-domain term extraction benefit from cross-lingual transfer? In *International Conference on Discovery Science* (pp. 363–378). Springer.

Tran, H. T. H., Martinc, M., Doucet, A., & Pollak, S. (2022). A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction. In *Slovenian Conference on Language Technologies and Digital Humanities*.

Tran, H. T. H., Martinc, M., Pelicon, A., Doucet, A., & Pollak, S. (2022). Ensembling transformers for cross-domain automatic term extraction. In *International Conference on Asian Digital Libraries* (pp. 90–100). Springer.

Vintar, Š. (2004). Comparative evaluation of c-value in the treatment of nested terms. In *Workshop Description* (pp. 54–57).

Vintar, S. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *International Journal of Theoretical and Applied Issues in Specialized Communication, 16*(2), 141–158.

Wolf, P., Bernardi, U., Federmann, C., & Hunsicker, S. (2011). From statistical term extraction to hybrid machine translation. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Zhang, Z., Gao, J., & Ciravegna, F. (2018). Semre-rank: Improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. *ACM Transactions on Knowledge Discovery from Data (TKDD), 12*(5), 1–41.

Chapter 5

Conclusions

This thesis presents various novel strategies for terminology extraction and alignment of terms between languages. We started by focusing on extraction and alignment from bilingual parallel corpora due to their prevalence, in the form of translation memories, in the translation industry. We then expanded our research in two directions: to terminology extraction from monolingual corpora, since having the ability to generate valid terms from texts is highly advantageous for the translation industry, and to aligning terminology from comparable corpora, partly due to their prominence among open-source data, and because it is a natural next step after generating monolingual terminology. Initial approaches were based on improving existing traditional methodologies and combining them with other machine learning strategies, such as evolutionary algorithms and supervised machine learning. After the emergence of neural approaches, we switched our focus from traditional methods to novel neural ones by incorporating pre-trained and contextual word embeddings, transformers and sequence labelling into our work. This chapter is organized as follows: in Section 5.1 we summarize the scientific contributions of the research, in Section 5.2 we discuss the strengths and weaknesses of the proposed approaches and in Section 5.3, we provide several avenues for future work.

5.1 Summary of Scientific Contributions

We have proposed novel approaches for various aspects of terminology extraction and alignment, to address the three goals defined in Section 1.3 of this thesis.

Initial research was related to goal G1 by attempting *to improve the performance of existing terminology alignment algorithms on parallel corpora from the translation industry.* We reimplemented several existing terminology alignment approaches and developed a novel approach using co-occurence statistics and phrase tables generated by an unsupervised model for joint phrase alignment and extraction using nonparametric Bayesian methods and inversion transduction grammars. The novel approach as well as existing approaches were then combined together via an evolutionary algorithm to determine the best combination of weights for optimal alignment result. This resulted in improved performance over individual approaches which means that we were also able to confirm hypothesis H1.1 stating that *terminology alignment performance can be significantly improved if we focus on the sub-sentence level via the construction of phrase tables and application of co-occurrence heuristics on the aligned phrases* and hypothesis H1.2 stating that *improved term alignment can be achieved by an ensemble approach, and that the complex problem of finding the optimal combination of different metrics can be effectively solved using evolutionary algorithms.*

Next we focused on goal G2 by attempting *to improve the performance of existing termi-*

nology alignment algorithms on comparable corpora. During the course of reimplementation and adaptation of an existing machine learning approach to terminology alignment, we developed several novel machine learning features based on word similarity across languages (i.e. cognates). Once aspect that we initially has problems with was that the classification algorithm returned modest precision values and high recall values which effectively meant that almost all candidates were classified as true terms. We applied dataset filtering techniques and made sure that the true/false ratio of the training and test sets were similar which resulted in higher precision and lower recall. Finally, we discovered that if we apply novel features based on word similarity across languages (i.e. cognates) we are able to boost recall while keeping precision at respectable levels. In later work, we also proposed additional novel features based on pre-trained cross-lingual word embeddings and contextual word and sentence embeddings. In doing so, we were able to confirm hypothesis H2 stating that *improved term alignment can be achieved by enhancing traditional terminology alignment features based on word alignment and cognate scores with novel features constructed by cross-lingual word embeddings.*

Finally, we turned to goal G3 by attempting *to improve the performance of existing terminology extraction algorithms from specialized corpora.* One avenue we explored was treating terminology extraction as a binary classification problem — for each candidate term we generated three types of features (statistical, linguistic and contextual) and evaluated the performance on a dedicated terminology extraction dataset (which we collaborated on as part of the the national language technology project RSDO). The statistical features were based on previous work by Vintar (2010) whereas linguistic and contextual features were novel. For linguistic features, we were inspired by traditional approaches that define POS patterns and extract term candidates that match those patterns. Instead of manually creating fixed patterns, we generate binary features for each Universal Dependency (UD) tag in various positions (e.g., at the start of the term, at the end of the term etc.) and allow the classifier to assess their significance. Finally, contextual features are generated using the ELMo contextual embeddings by calculating the average term embedding in a specialised corpus and comparing it with the average term embedding in a general language corpus, by comparing it to the average term embedding of all terms in the specialised corpus, by comparing it to a typical representative term of a specialised corpus and by measuring the standard deviation of all term occurrences in the specialised corpus. We were able to exceed the performance of existing terminology extraction algorithms for Slovenian by a large margin and our approach proved particularly useful for low-frequency terms where traditional frequency-based approaches are weak. The results of this approach mean that we were able to confirm hypothesis H3.1 stating that *applying contextual information from deep neural network models by exploiting the contextual differences in the domain and general corpora can help to improve terminology extraction performance.* Finally, as transformer-based models in sequence labelling settings showed promising results also for terminology extraction (Lang et al., 2021), we adapted it to the Slovenian setting and tested a novel labelling scheme for nested terms and achieved state-of-the-art results which confirmed hypothesis H3.2 stating that *by using richer contextual information, these embeddings can better capture subtle meanings and domain-specific details leading to more accurate detection of terms in text.*

5.2 Discussing the Strength and Weaknesses of the Developed Approaches

In this section, we will discuss the strengths of the developed approaches and critically evaluate their weaknesses. We have proposed novel approaches in the area of terminology

alignment and terminology extraction by combining traditional linguistic and statistical approaches with advanced machine learning and neural network-based natural language processing techniques. Their performance was measured in terms of standard measures of F_1, precision and recall[1], but we will also discuss other important aspects. Since NLP research is often focused on a single language and being able to adapt the method for other languages is a key advantage, we will discuss whether our approaches can be easily adapted for other languages. We will discuss the intepretability of the developed approaches, i.e. whether it is possible to get in insight into the functioning of the method for easier debugging and transfer of findings to other areas and methods. Finally, we will also look at whether the developed approaches can be easily replicated (i.e. their replicability).

5.2.1 Bilingual terminology extraction and alignment from parallel corpora

The bilingual terminology extraction and alignment method from parallel corpora, described in Chapter 2, was evaluated using manual evaluation. The method was developed for a translation industry client who was primarily interested in precision and they were able to provide a domain expert to evaluate the results. All proposed bilingual term pairs were assigned a rank by the algorithm and were manually evaluated by the domain expert and we then used a Top N approach to calculate the precision, i.e. the higher the number of correct term pairs in the top N places, the better the precision. Performance is one of the strengths of this method, since we achieved a precision of 0.960 or more in three different domains on the top 400 term pairs using the evolutionary algorithm approach. The method was developed for and tested on Slovenian data, but can be adapted to other languages with the exception of the monolingual extraction component, which requires a language-specific list of POS patterns. In terms of interpretability, some elements are not entirely straightforward (e.g., the phrase-table-based alignment and the evolutionary algorithm) but they are still more interpretable than more novel neural approaches. Finally, the approach cannot be replicated because it was developed for an industrial client and is not publicly available.

5.2.2 Bilingual terminology alignment from comparable corpora

The bilingual terminology alignment method from comparable corpora, described in Chapter 3, was evaluated with standard evaluation metrics of precision, recall and F_1 score. The highest F_1 score, achieved on the Slovenian dataset, was 0.67, but we must note the specific evaluation setting: the dataset was not created for bilingual terminology alignment in mind where one would expect that every possible valid pair is annotated. Instead, we used Eurovoc (Steinberger et al., 2002) which is a thesaurus of EU related terminology. This means that there our classifier may well have predicted proper term pairs, but since they were not part of Eurovoc, they were counted as false positives. The approach in Repar et al. (2020) was additionally manually evaluated on two domains (finance and information technology) to replicate the evaluation experiments in the original paper and on a third domain (karstology) with a specific focus on cognate term pairs, which was one of the novelties of the paper and which showed improved performance on cognate terms. The method was developed for and tested on three language pairs: English-Slovenian, English-French and English-Dutch. The version from Repar et al. (2020) requires a large bilingual corpus to generate the Giza++ dictionary, but the adaptation in Repar et al. (2021) only needs a small seed dictionary to generate the cross-lingual embeddings needed to find alignments,

[1]With the exception of the approach in Repar et al. (2019) where we used a Top N evaluation approach combined with manual evaluation as we did not have a dataset that would allow us to evaluate recall.

which means that the method can be adapted to other languages fairly easily. In terms of interpretability, the core part of the method (i.e. classification) is easily interpretable, but individual features are derived from word embeddings which makes them less interpretable. The method can be replicated: the code is available publicly and a section of Repar et al. (2020) is dedicated to an experiment on the ease of replicability with all three participating students being able to reproduce the results in less than 2 hours.

5.2.3 A machine learning approach to monolingual terminology extraction

The machine learning approach to monolingual terminology extraction, described in Chapter 4, was evaluated with standard evaluation metrics of precision, recall and F_1 score. We achieved F_1 score values between 0.530 and 0.594 on different test domains which meant that we significantly improved on the previous state-of-the-art approach on the Slovenian language and that we achieve performance comparable to other approaches on other languages as reported in the TermEval shared task (Rigouts Terryn, Hoste, Drouin, et al., 2020). However, these results were exceeded by sequence labeling approaches (e.g., Lang et al. (2021)) which were developed concurrently with our research. Despite not producing state-of-the-art results, we believe that there are aspects of this research can provide better insights into the process of terminology extraction. In terms of interpretability, our machine learning approach is based on linguistic intuition and confirms the basic definition of a term from the ISO 1087 standard on "Terminology work and terminology science" which is that a term is a "verbal designation of a general concept in a specific subject field". The contextual features in our study are designed using the premise that a term appears in different contexts when comparing domain corpora and general language corpora. In addition, we also show an additional characteristic of terms (via the *elmoStDev* feature) which is that even inside a domain-specific corpus, terms tend to appear in the same context. Another positive aspect of our approach compared to sequence labeling is that it works with lemmas as opposed to word forms. While this may not be a large issue for languages with few word forms, such as English where sequence labeling was first evaluated, it is far less intuitive to use word forms on languages with a rich morphology, such as Slovenian. Note that this also means that the results of our approach and the sequence labeling approach in H. T. H. Tran, Martinc, Doucet, et al. (2022) are not directly comparable — for example, in the biomechanics domain, we have a total of 1596 unique lemma terms, whereas H. T. H. Tran, Martinc, Doucet, et al. (2022) use a total of 2319 unique word forms[2]. On the other hand, our approach also has some drawbacks. In terms of adaptability to other languages, while it improves upon other legacy approaches based on POS patterns by employing shallow filter rules, one still requires language-specific knowledge to generate these rules. The approach is also dependent on language-specific resources, in particular ELMo contextual embeddings for a general language corpora. While these corpora are available for many languages, the calculation of contextual embeddings is resource intensive and takes a long time.

[2]We tried comparing the results of the two approaches by assuming that for each lemma term predicted by our model, we also predicted all the word forms associated with it. By doing so, we were able to achieve F_1 scores of around or a bit above 0.6 — still not better than sequence labeling but perhaps a more accurate comparison

5.2.4 A sequence labeling approach to monolingual terminology extraction

The sequence labeling method achieves the best results overall, with F_1 scores often above 0.6 and even reaching over 0.7 on the RSDO5 corpus, making it the most promising for further research. However, it has some limitations. This method focuses only on individual sentences and ignores broader patterns in the corpus, like how often terms appear or their context. Adding these broader factors could improve results. It is also a "black box" method, meaning it doesn't rely on linguistic understanding, which makes it hard to explain why certain phrases are chosen as terms. Another issue is that it works with word forms, not lemmas. This is less of a problem for languages like English, where word forms and lemmas are often similar (e.g., "house" and "houses" both have the lemma "house"). However, for languages with many word forms, like Slovenian, this difference is significant because such languages have a lot more variation in word endings (e.g., 6 cases, plus dual and plural forms).

5.2.5 Practical usability of the developed approaches in the translation industry

With the exception of the first approach, all developed methods utilize publicly available gold standard datasets and standard evaluation metrics. While these datasets are valuable for benchmarking, they are typically constructed as static, comprehensive lists of terms. In practice, however, many domain-specific terms are already familiar to translators. For instance, although the term "heart" may appear in a gold standard dataset on heart failure, it provides little practical utility to a translator who already knows the term. A truly useful tool would be capable of highlighting only those terms that are unfamiliar to the translator—a concept further discussed in the context of termbanks in Section 5.3. Moreover, the benefits of terminology extraction and alignment tools extend beyond the translation phase. When working with new clients, translators and translation companies can leverage these tools to automatically generate term lists and glossaries from a client's existing multilingual documentation. This not only accelerates the onboarding process but also helps identify inconsistencies and errors in the client's materials, ultimately improving translation quality and consistency. While standard evaluation metrics such as recall are commonly used in research, they are less practical in the translation industry. Recall measures how many relevant terms are identified by a tool, but defining what constitutes a term can be inherently subjective and domain-dependent. Certain words may be considered terms in one context but not in another, and translators may already be familiar with many extracted terms, rendering recall-based evaluations less meaningful. Instead, practical usability in the industry is better assessed by how effectively a tool surfaces useful and unknown terms to the translator rather than by maximizing the number of terms extracted.

5.3 Further Work

In the area of monolingual terminology extraction, we would like to investigate how to merge the best aspects of the machine learning and sequence labeling approaches to monolingual terminology extraction. Sequence labeling, given its superior performance, is the obvious baseline to work with, but there are several individual aspects of the machine learning approach, particularly those related to term context in a given corpus, that could help further improve the algorithm. Another possible avenue to explore would be to replace ELMo embeddings with another model (e.g., BERT) which general performs better

in NLP tasks. In terminology alignment, one area to explore is using co-occurence in aligned sentence pairs (either from a smaller sub-corpus of aligned pairs or from synthetically aligned data) as an additional classifier feature on top of other features based on comparable corpora.

However, given the recent emergence of large language models (LLMs) as the dominant technology, future research will increasingly focus on their potential for extracting and aligning terminology. LLMs have gained widespread traction in research and hold significant promise for enhancing terminology extraction and alignment, as well as improving translation quality. As in many other areas of NLP, LLMs are poised to shape the future of terminology work thanks to their ease of domain adaptation (e.g. via in-context learning (ICL)), ability to process large context windows, and natural language interfaces enabled by prompting. This shift does not signal the end of terminology work; rather, it will evolve to integrate and leverage the capabilities of LLMs even more heavily.

One approach for leveraging LLMs in translation is in-context learning (ICL), combined with prompt engineering techniques, as shown in studies like the one by H. T. H. Tran et al. (2024). This approach involves providing an LLM with a few examples of the desired output or clearly explaining the task to it. Professional translators are also beginning to adopt LLMs in ways they didn't with other language technologies, mainly because LLMs are easy to use and readily accessible. In comparison, creating a specialized terminology extraction model using machine learning or sequence labeling is often labor-intensive, requiring specific data sets, software, or a user interface. By contrast, LLMs are easy to access via chat interfaces, enabling anyone to experiment with prompts and see immediate results. Translators bring a unique perspective here. While they may not have formal training in prompt engineering or computing, they have deep expertise in their field, such as legal translation, and are skilled at communicating in multiple languages. Their insights could prove valuable for optimizing prompts to achieve the best results in terminology extraction and alignment. These insights could, in turn, contribute to advancements in the field of prompt engineering itself.

A potential use of LLMs in translation work involves a practice that has often been overlooked: the creation of "termbanks." Many translators build these termbanks as they work, compiling terms, standard phrases, and helpful expressions that can streamline future translations. Unlike gold standard terminology datasets, these termbanks aren't comprehensive—they don't aim to cover all terms from a document or corpus. This is often by design; terms that are easy to translate, even if common in a specific field (like "heart" in medical contexts), are excluded to keep the interface cleaner and more useful in translation tools. One promising direction would be to fine-tune an LLM using various termbanks or use them, along with other valuable resources, in a retrieval augmented generation (RAG) scenario, so that it could generate relevant termbanks for new documents or sets of documents before translation begins or during the translation process itself. This could address a limitation in many existing terminology extraction tools, which often produce a broad set of valid terms, including some that are simple to translate and therefore less useful to translators. Fine-tuned LLMs or RAG workflows could potentially offer more targeted termbanks, highlighting new or challenging terms and phrases that would be more practical for translators.

Another reason why a RAG approach could be effective is the ability to use high-quality data as the foundation for terminology extraction and aligment. LLM training data often contains noise, such as outputs from machine translation (MT) engines or previously generated LLM text, which can lead to suboptimal term suggestions. For example, the English term fan coil unit (FCU), used in split HVAC systems, translates properly to Slovenian as "ventilatorski konvektor". However, if you input "fan coil unit" into an MT engine or

even ChatGPT, the best result you can usually get is "ventilatorska konvektorska enota". While this translation is not incorrect, it is unnatural because the term "enota" ("unit") is implied in the word "konvektor" through its ending ("or"). Such nuances in terminology are critical for maintaining naturalness and precision in specialized translations.

Finally, large foundational models are trained on data in multiple languages, giving them the ability to understand and work across different languages. This capability makes them promising tools for identifying terms in one language and finding their equivalent in another, especially within specific domains. Unlike traditional methods, which often treat terminology extraction and alignment as two separate steps, foundational models can potentially handle both tasks in a single, unified process. We plan to explore how LLMs can be leveraged to develop a streamlined multilingual terminology extraction and alignment workflow. By directly linking term extraction with cross-lingual alignment, we could eliminate the need for separate tools or sequential processes, reducing complexity and improving efficiency by having the model to simultaneously identify domain-specific terms in a source language and generate aligned terms in the target language.

References

Adjali, O., Morin, E., Sharoff, S., Rapp, R., & Zweigenbaum, P. (2022). Overview of the 2022 bucc shared task: Bilingual term alignment in comparable specialized corpora. *BUCC, 15th Workshop on Building and Using Comparable Corpora*, 67–76.

Adjali, O., Morin, E., & Zweigenbaum, P. (2022). Building comparable corpora for assessing multi-word term alignment. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 3103–3112). European Language Resources Association. https://aclanthology.org/2022.lrec-1.332/

Aker, A., Paramita, M., & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *1*, 402–411.

Amjadian, E., Inkpen, D., Paribakht, T., & Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, 2–11.

Artetxe, M., Labaka, G., & Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Artetxe, M., Labaka, G., & Agirre, E. (2019). Bilingual lexicon induction through unsupervised machine translation. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5002–5007). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1494

Baisa, V., Ulipová, B., & Cukr, M. (2015). Bilingual terminology extraction in sketch engine. *RASLAN*, 61–67.

Baldwin, T., & Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 24–31.

Banerjee, S., Chakravarthi, B. R., & McCrae, J. P. (2024). Large language models for few-shot automatic term extraction. In A. Rapp, L. Di Caro, F. Meziane, & V. Sugumaran (Eds.), *Natural language processing and information systems* (pp. 137–150). Springer Nature Switzerland.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Cao, Y., & Li, H. (2002). Base noun phrase translation using web data and the em algorithm. *COLING 2002: The 19th International Conference on Computational Linguistics*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. https://arxiv.org/abs/1710.04087

Conrado, M., Felippo, A., Pardo, T., & Rezende, S. (2014). A survey of automatic term extraction for brazilian portuguese. *Journal of the Brazilian Computer Society*, *20*, 12. https://doi.org/10.1186/1678-4804-20-12

Daille, B., Gaussier, É., & Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Daille, B., & Morin, E. (2005). French-English terminology extraction from comparable corpora. *Second International Joint Conference on Natural Language Processing: Full Papers*. https://doi.org/10.1007/11562214_62

De Rainville, F.-M., Fortin, F.-A., Gardner, M.-A., Parizeau, M., & Gagné, C. (2012). Deap: A python framework for evolutionary algorithms. *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*, 85–92. https://doi.org/10.1145/2330784.2330799

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Erjavec, T., Fišer, D., Ljubešić, N., Arhar Holdt, Š., Bren, U., Robnik-Šikonja, M., & Udovič, B. (2018). Terminology identification dataset KAS-term 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1198

Fišer, D., Suchomel, V., & Jakubı´cek, M. (2016). Terminology extraction for academic Slovene using Sketch Engine. *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, 135–141.

Foo, J. (2012). *Computational terminology: Exploring bilingual and monolingual term extraction* (Doctoral dissertation). Linköping University Electronic Press.

Foo, J., & Merkel, M. (2010). Using machine learning to perform automatic term recognition. *LREC 2010 Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods*, 49–54.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/ NC-value method. *International Journal on Digital Libraries*, *3*(2), 115–130. https://doi.org/10.1007/3-540-49653-X_35

Gao, Y., & Yuan, Y. (2019). Feature-less End-to-end Nested Term extraction. *CCF International Conference on Natural Language Processing and Chinese Computing*, 607–616.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 1–6.

Giguere, J. (2023). Leveraging large language models to extract terminology. In R. L. Gutiérrez, A. Pareja, & R. Mitkov (Eds.), *Proceedings of the first workshop on nlp tools and resources for translation and interpreting applications* (pp. 57–60). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.nlp4tia-1.9

Haque, R., Penkale, S., & Way, A. (2014). Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. *Proceedings of the 4th international workshop on computational terminology (Computerm)*, 42–51.

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2022). Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 648–662. https://aclanthology.org/2022.lrec-1.68

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2020). TermEval 2020: TALN-LS2N system for automatic term extraction. *Proceedings of the 6th International Workshop on Computational Terminology*, 95–100.

Hazem, A., & Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 685–693.

Jemec Tomažin, M., Trojar, M., Žagar, M., Atelšek, S., Fajfar, T., & Erjavec, T. (2021). Corpus of term-annotated texts RSDO5 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1400

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition. a review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *3*(2), 259–289.

Kageura, K., & Umino, B. (2001). Methods of automatic term recognition — a review —. *Terminology*, *3*. https://doi.org/10.1075/term.3.2.03kag

Karan, M., Šnajder, J., & Bašic, B. D. (2012). Evaluation of classification algorithms and features for collocation extraction in Croatian. *Proceedings of the 12th Language Resources and Evaluation Conference*, 657–662.

Khan, M. T., Ma, Y., & Kim, J.-j. (2016). Term Ranker: A Graph-Based Re-Ranking Approach. *FLAIRS Conference*, 310–315.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 177–180.

Krek, S., Holdt, Š. A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: The reference corpus of written standard Slovene. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3340–3345.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., & Stüker, S. (2018a). Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. *INTERSPEECH*, 2072–2076.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., & Stüker, S. (2018b). Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks, 2072–2076. https://doi.org/10.21437/Interspeech.2018-2017

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. *31st Annual Meeting of the Association for Computational Linguistics*, 17–22.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270.

Lang, C., Wachowiak, L., Heinisch, B., & Gromann, D. (2021). Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3607–3620. https://doi.org/10.18653/v1/2021.findings-acl.316

Ljubešić, N., & Dobrovoljc, K. (2019). What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 29–34.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). Kas-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. *International Conference on Text, Speech, and Dialogue*, 115–126.

Logar, N., Erjavec, T., Krek, S., Grčar, M., & Holozan, P. (2013). Written corpus ccGigafida 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1035

Macken, L., Lefever, E., & Hoste, V. (2013). Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *19*(1), 1–30.

Martinc, M., Škrlj, B., & Pollak, S. (2022). Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, *28*(4), 409–448. https://doi.org/10.1017/S1351324921000127

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Meex, B., & Straub, D. (2016). How to bridge the gap between translators and technical communicators? the importance of sharing knowledge to improve the localization process. *The journal of internationalization and localization*, *3*(2), 133–151.

Meyers, A. L., He, Y., Glass, Z., Ortega, J., Liao, S., Grieve-Smith, A., Grishman, R., & Babko-Malaya, O. (2018). The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores. *Frontiers in Research Metrics and Analytics*, *3*, 19.

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*.

Mustafa, W., El Hadi, W., Timimi, I., Dabbadie, M., Choukri, K., Hamon, O., & Chiao, Y.-C. (2006). Terminological resources acquisition tools: Toward a user-oriented evaluation model.

Nassirudin, M., & Purwarianti, A. (2015). Indonesian-japanese term extraction from bilingual corpora using machine learning. *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 111–116.

Neubig, G., Watanabe, T., Sumita, E., Mori, S., & Kawahara, T. (2011). An unsupervised model for joint phrase alignment and extraction. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 632–641.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1*

(long papers) (pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Pinnis, M., Ljubešic, N., Stefanescu, D., Skadina, I., Tadic, M., & Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. *Proceedings of the 10th Conference on Terminology and Knowledge Engineering*, 20–21.

Požár, B., Tauchmanová, K., Neumannová, K., Kvapilíková, I., & Bojar, O. (2022). CUNI submission to the BUCC 2022 shared task on bilingual term alignment. In R. Rapp, P. Zweigenbaum, & S. Sharoff (Eds.), *Proceedings of the bucc workshop within lrec 2022* (pp. 43–49). European Language Resources Association. https://aclanthology.org/2022.bucc-1.6/

Reimers, N., & Gurevych, I. (2019a). Alternative weighting schemes for ELMo embeddings. *arXiv preprint arXiv:1904.02954*.

Reimers, N., & Gurevych, I. (2019b). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, *abs/1908.10084*. http://arxiv.org/abs/1908.10084

Repar, A., Pollak, S., Ulčar, M., & Koloski, B. (2022). Fusion of linguistic, neural and sentence-transformer features for improved term alignment. In R. Rapp, P. Zweigenbaum, & S. Sharoff (Eds.), *Proceedings of the bucc workshop within lrec 2022* (pp. 61–66). European Language Resources Association. https://aclanthology.org/2022.bucc-1.9

Repar, A., Martinc, M., & Pollak, S. (2020). Reproduction, replication, analysis and adaptation of a term alignment approach. *Language Resources and Evaluation*, *54*(3), 767–800.

Repar, A., Martinc, M., Ulcar, M., & Pollak, S. (2021). Word-embedding based bilingual terminology alignment. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, 98.

Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., & Pollak, S. (2019). TermEnsembler: An Ensemble Learning Approach to Bilingual Term Extraction and Alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *25*(1), 93–120.

Repar, A., & Shumakov, A. (2021). Aligning Estonian and Russian news industry keywords with the help of subtitle translations and an environmental thesaurus. In H. Toivonen & M. Boggia (Eds.), *Proceedings of the eacl hackashop on news media content analysis and automated report generation* (pp. 71–75). Association for Computational Linguistics. https://aclanthology.org/2021.hackashop-1.10

Rigouts Terryn, A., Hoste, V., Drouin, P., & Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, 85–94.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2020a). In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation*, *54*(2), 385–418.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2020b). In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, *54*(2), 385–418.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2021a). HAMLET: Hybrid adaptable machine learning approach to extract terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *27*(2), 254–293.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2021b). Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology. Interna-*

tional Journal of Theoretical and Applied Issues in Specialized Communication, *28*. https://doi.org/10.1075/term.21010.rig

Setha, I., & Aliane, H. (2023). Bilingual terminology alignment using contextualized embeddings. In A. H. Haddad, A. R. Terryn, R. Mitkov, R. Rapp, P. Zweigenbaum, & S. Sharoff (Eds.), *Proceedings of the workshop on computational terminology in nlp and translation studies (contents) incorporating the 16th workshop on building and using comparable corpora (bucc)* (pp. 1–8). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.contents-1.1/

Shi, H., Zettlemoyer, L., & Wang, S. I. (2021). Bilingual lexicon induction via unsupervised bitext construction and word alignment. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 813–826). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.67

Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, 101–121.

Thompson, B., Dhaliwal, M., Frisch, P., Domhan, T., & Federico, M. (2024). A shocking amount of the web is machine translated: Insights from multi-way parallelism. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics acl 2024* (pp. 1763–1775). Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.103

Tran, H., Martinc, M., Repar, A., Ljubešić, N., Doucet, A., & Pollak, S. (2024). Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling? *Machine Learning*, *113*, 1–30. https://doi.org/10.1007/s10994-023-06506-7

Tran, H., Martinc, M., Repar, A., Doucet, A., & Pollak, S. (2022). A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction. *Proceedings of the Conference on Language Technologies and Digital Humanities*.

Tran, H. T. H., Doucet, A., Sidere, N., Moreno, J., & Pollak, S. (2021). Named Entity Recognition Architecture Combining Contextual and Global Features. In H.-R. Ke, C. S. Lee, & K. Sugiyama (Eds.), *Towards Open and Trustworthy Digital Societies. 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings* (pp. 264–276). Springer. https://doi.org/10.1007/978-3-030-91669-5_21

Tran, H. T. H., González-Gallardo, C.-E., Delaunay, J., Doucet, A., & Pollak, S. (2024). Is prompting what term extraction needs? *International Conference on Text, Speech, and Dialogue*, 17–29.

Tran, H. T. H., Martinc, M., Caporusso, J., Doucet, A., & Pollak, S. (2023). The recent advances in automatic term extraction: A survey. https://arxiv.org/abs/2301.06767

Tran, H. T. H., Martinc, M., Doucet, A., & Pollak, S. (2022). Can cross-domain term extraction benefit from cross-lingual transfer? *Discovery Science: 25th International Conference, DS 2022, Montpellier, France, October 10–12, 2022, Proceedings*, 363–378. https://doi.org/10.1007/978-3-031-18840-4_26

Tran, H. T. H., Martinc, M., Pelicon, A., Doucet, A., & Pollak, S. (2022). Ensembling transformers for cross-domain automatic term extraction. In *From born-physical to born-virtual: Augmenting intelligence in digital libraries* (pp. 90–100). Springer International Publishing. https://doi.org/10.1007/978-3-031-21756-2_7

Ulčar, M. (2019). ELMo embeddings models for seven languages [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1277

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., & Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. https://arxiv.org/abs/2107.10614

Vintar, S. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *16*(2), 141–158.

Wang, R., Liu, W., & McDonald, C. (2016). Featureless Domain-Specific Term Extraction with Minimal Labelled Data. *Proceedings of the Australasian Language Technology Association Workshop 2016*, 103–112.

Zadeh, B., & Handschuh, S. (2014). Evaluation of technology term recognition with random indexing. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

Zhang, Z., Gao, J., & Ciravegna, F. (2017). SemRe-Rank: Incorporating Semantic Relatedness to Improve Automatic Term Extraction Using Personalized PageRank. *arXiv preprint arXiv:1711.03373*.

Bibliography

Publications Related to the Thesis

Journal Articles

Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., & Pollak, S. (2019). TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment. Terminology, 25(1), 93-120.

Repar, A., Martinc, M., & Pollak, S. (2020). Reproduction, replication, analysis and adaptation of a term alignment approach. *Language resources and evaluation*, 54(3), 767-800.

Tran, H., Martinc, M., Repar, A., Ljubešić, N., Doucet, A., & Pollak, S. (2024). Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling?. Machine Learning, 113, 1-30.

Conference and Workshop Papers

Repar, A., Martinc, M., Ulčar, M., & Pollak, S. (2021). Word-embedding based bilingual terminology alignment. Proceedings of Electronic lexicography in the 21st century: Post-editing lexicography, 408-417.

Repar, A., Pollak, S., Ulčar, M., & Koloski, B. (2022). Fusion of linguistic, neural and sentence-transformer features for improved term alignment. Proceedings of the BUCC Workshop within LREC 2022, 61-66.

Repar, A., & Shumakov, A. (2021). Aligning Estonian and Russian news industry keywords with the help of subtitle translations and an environmental thesaurus. Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, 71–75.

Tran, H., Martinc, M., Repar, A., Doucet, A., & Pollak, S. (2022). A Transformer-based sequence-labeling approach to the Slovenian cross-domain automatic term extraction. *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia, 196-204.

Other Publications

Repar, A., Martinc, M., & Pollak, S. (2018). Machine learning approach to bilingual terminology alignment : reimplementation and adaptation. *Proceedings of the 4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*, Miyazaki, Japan, 1-8.

Pollak, S., Repar., A., Martinc, M., & Podpečan, V. (2019). Karst exploration : extracting terms and definitions from Karst domain corpus. *Proceedings of the eLex 2019 Conference*, Sintra, Portugal, 934-956.

Repar, A., Martinc, M., Žnidaršič, M., & Pollak, S. (2018). BISLON: BISociative SLOgaN generation based on stylistic literary devices. *Proceedings of the Ninth International Conference on Computational Creativity*, Salamanca, Spain, 248-255.

Vintar, Š., & Repar, A. (2022). Human evaluation of machine translations by semi-professionals : lessons learnt. *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia, 220-226.

Váradi, T., Nyéki, B., Krek, S., Repar, A. et al. (2022) Introducing the CURLICAT Corpora : seven-language domain specific annotated corpora from curated sources. *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France, 100-108.

Váradi, Ta., Tadić, M., Krek, S., Repar, A. et al. (2022) Curated Multilingual Language Resources for CEF AT (CURLICAT) : overall view. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, Ghent, Belgium, 341-342.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0 : the reference corpus of written standard Slovene. *Proceedings of the 12th International Conference on Language Resources and Evaluation*, Marseille, France, 3340-3345.

Váradi, T., Krek, S., Repar, A., Rihtar, M., Brank, J. et al. (2020). The MARCELL legislative corpus. *Proceedings of the 12th International Conference on Language Resources and Evaluation*, Marseille, France, 3761-3768.

Škrlj, B., Repar, A., & Pollak, S. (2019). RaKUn: rank-based keyword extraction via unsupervised learning and meta vertex aggregation. *Proceedings of the 7th International Conference on Statistical Language and Speech Processing*, Ljubljana, Slovenia, 311-323.

Repar, A., & Pollak, S. (2017) Good examples for terminology databases in translation industry. *Electronic lexicography in the 21st century : proceedings of eLex 2017 Conference*, Leiden, The Netherlands, 650-661.

Biography

Andraž Repar was born in April 27, 1985 in Ljubljana, Slovenia. He studied translation studies at the Faculty of Arts at the University of Ljubljana, where he defended his bachelor thesis "A glossary of heating, ventilation and air-conditioning terms" in 2010. After receiving his diploma, he started working at the translation company Iolar first as a translator and later as a reviser and quality manager. During his time at the company, he became familiar with workings of the translation industry and eventually recognized terminology as the key factor affecting translation quality. In 2018, he formed his own translation company with three other colleagues where he is now responsible for business development and overall process improvement.

In 2016, he enrolled in the PhD programme "Information and Communication Technologies" at the Jožef Stefan International Postgraduate School under the supervision of Assist. Prof. Dr. Senja Pollak.

His research is focused on terminology extraction and alignment and combines linguistic aspects with natural language processing and machine learning. He authored a number of articles on terminology extraction and alignment, as well as on several related topics, such as keyword extraction and good examples. He was involved in several European Union's Horizon 2020 projects and Connecting Europe Facility projects, such as EMBEDDIA, ELEXIS, MARCELL and CURLICAT, and national projects TermFrame and RSDO.