Can Cross-domain Term Extraction Benefit from Cross-lingual Transfer?

Tran Thi Hong Hanh, Matej Martinc, Antoine Doucet, Senja Pollak

Can Cross-domain Term Extraction Benefit from Cross-lingual Transfer?

Hanh Thi Hong TRAN$^{1,2,3[0000-0002-5993-1630]}$, Matej MARTINC$^{1[0000-0002-7384-8112]}$, Antoine DOUCET$^{3[0000-0001-6160-3356]}$, and Senja POLLAK$^{2[0000-0002-4380-0863]}$

[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

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 experiment with XLM-RoBERTa to evaluate the abilities of cross-lingual and multilingual versus monolingual learning in the cross-domain ATE task. The experiments are conducted on the ACTER corpus covering four domains (Corruption, Wind energy, Equitation, and Heart failure) and three languages (English, French, and Dutch) and on the RSDO5 Slovenian corpus, covering four additional domains (Biomechanics, Chemistry, Veterinary, and Linguistics). Regarding the ACTER test set, the cross-lingual and multilingual models boost the performance in F1-score by up to 5% if the term extraction task excludes the extraction of named entity terms (ANN version) and 3% if including them (NES version) compared to the monolingual setting. By adding an extra Slovenian corpus into the training set, the multilingual model demonstrates a significant improvement in terms of Recall, which, on average, increases by 18% in the ANN version and 13% in the NES version compared with the monolingual setting. Furthermore, our methods defeat state-of-the-art (SOTA) approaches with approximately 2% higher F1-score on average for the ANN version in English and Dutch, and the NES version in French. Regarding the RSDO5 test set, our monolingual approach proves to have consistent performance across all the train-validation-test combinations, achieving an F1-score above 61%. These results are a good indication of the potential in cross-lingual and multilingual language models not only for term extraction but also for other downstream tasks. Our code is publicly available at https://github.com/honghanhh/ate-2022.

Keywords: Term extraction · XLM-RoBERTa · Sequence labeling · Cross-lingual · Cross-domain.

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 term dictionaries [21]). Moreover, terms can also support and improve several complex downstream natural language processing (NLP) tasks, such as topic detection [6], information retrieval [22], machine translation [36], etc. Automatic term extraction (ATE) was born to ease the time and effort needed to manually identify terms from domain-specific corpora.

The TermEval 2020 shared task on monolingual ATE, organized as part of the CompuTerm workshop [30], presented one of the first opportunities to systematically study and compare various ATE systems with the introduction of a new annotated corpus that covers four domains in three languages: The Annotated Corpora for Term Extraction Research (ACTER) dataset [30,31]. While the workshop was an important 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 and replicability as the open-sourced code is often not available.

Inspired by the success of Transformer-based models in the TermEval 2020 competition [11] and the rise of cross-lingual learning [19], we propose to explore the performance of the multilingual XLM-RoBERTa pretrained model [3] in a multilingual setting, and in a cross-lingual setting, where the model is fine-tuned on several languages and tested on a new unseen language. We model the ATE as a sequence-labeling task. Sequence-labeling approaches have been successfully applied to a range of similar NLP tasks, including Named Entity Recognition [18,34] and Keyword Extraction [25,16]. The experiments are conducted in the cross-domain setting on the ACTER dataset containing texts in four domains (Corruption, Wind energy, Equitation, and Heart failure) with three languages (English, French, and Dutch) and the RSDO5 corpus[4] [12] containing Slovenian texts from four domains (Biomechanics, Chemistry, Veterinary, and Linguistics).

The main contributions of this paper can be summarized as follows:

- We systematically evaluate the performance of XLM-RoBERTa language model on the cross-domain term extraction task on two datasets covering English, French, Dutch, and a less-resourced language, Slovenian.
- We compare the performance of cross- and multilingual toward monolingual approaches to determine the general applicability of multilingual language models for sequence labeling in both rich- and less-resourced languages, for which manually labeled training resources are and are not available.

This paper is organized as follows: Section 2 presents the related work. Next, we introduce the dataset, methodology, experimental details as well as evaluation metrics in Section 3. The results with error analysis are discussed in Section 4 and 5 before we conclude and present future works in Section 6.

[4] https://www.clarin.si/repository/xmlui/handle/11356/1470

2 Related Work

The history of ATE has its beginnings during the 1990s with research done by Damerau et al. [5] , Justeson et al. [14]. ATE systems usually employ the two-step procedure: (1) extracting a list of candidate terms; and (2) determining which candidate terms are correct using supervised or unsupervised techniques. We divide these techniques into the approaches based on (1) term characteristics and (2) machine learning and deep learning.

2.1 Approaches based on term characteristics

Traditional ATE approaches relied on linguistic knowledge and distinctive linguistic aspects of terms to extract possible candidates. Several NLP tools (e.g., tokenization, lemmatization, stemming, chunking, PoS tagging, etc.) 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 [1], Stanza [28]), the better the quality of linguistic ATE methods. More recently, several studies were proposed that preferred the statistical approach toward ATE. The most common statistical approach 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 [35], unithood [4] or C-value [9]. More popular statistical approaches take also into account the frequency of the term internal words compared to the term frequency (e.g., Mutual Information) to identify rare terms and remove frequent words. Many current systems still apply this approach's variation, most commonly in hybrid systems combining linguistic and statistical information [15,29].

2.2 Approaches based on machine learning and deep learning

Recently, advances in embeddings and deep neural networks have also influenced the field of term extraction. Several embeddings have been investigated for the task at hand, for example, uni-gram term representations constructed from a combination of local and global vectors [2], non-contextual [37], contextual [17] word embeddings, and the combination of both [10]. The first use of language models for the ATE task is documented in the TermEval 2020 [30] competition on the ACTER dataset, a collection of four domain-specific corpora in three languages (English, French, and Dutch). There, the winning approach on the Dutch corpus used pretrained GloVe word embeddings fed into a BiLSTM-based neural architecture. Meanwhile, the winning approach on the English corpus [11] 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, several variations of Transformer-based models have also been investigated (e.g., RoBERTa and CamemBERT have also been used in the TermEval 2020 [11] challenge). Further work inspired by TermEval 2020 includes the HAMLET [32], which proposes a hybrid adaptable machine learning approach that combines the linguistic and statistical clues to detect terms.

When it comes to more general related work applicable to ATE task, the research by Kucza et al. [17] was one of the first to propose to model term extraction as a sequence labeling task. Cross-lingual sequence labeling was, on the other hand, explored in Conneau et al. [3] and Lang et al. [19], who take advantage of XLM-RoBERTa, the model we also employ in this work, to compare three cross-lingual approaches, including a binary sequence classifier, a sequence classifier, and a token classifier on several sequence-labeling tasks. Finally, Lang et al. [19] further proposes to use a multilingual encoder-decoder denoising pretraining model called mBART [23] 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.

2.3 Approaches for Slovenian term extraction

When it comes to the ATE for Slovenian, and more generally to less-resourced languages, the research is still hindered by the lack of gold standard corpora and limited use of neural methods. The things are nevertheless slowly improving. For example, in recent years, Slovenian KAS corpus was compiled [7]. The release was quickly followed by another corpus designed for term extraction, the RSDO5 corpus that we use in our study [13]. Regarding the employment of ATE models for Slovenian, one of the first approaches was the statistical approach by Vintar et al. [35]. The SOTA was proposed by Ljubevsic et al. [24], where they extract the initial candidate terms using the CollTerm tool [27], a rule-based system employing a complex language-specific set of term patterns (e.g., POS tag,...) from the Slovenian SketchEngine module [8], followed by a machine learning classification approach with features representing statistical term extraction measures. Another recent approach by Repar et al. [29] 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 is the open-sourced code is often not available for most current benchmark systems, hindering their reproducibility (for Slovenian, only the code from Ljubevsic et al.'s method [24] is available). In our own work [33], we also implemented the Transformers-based sequence labeling approach that we extend in this study in a cross-lingual and multilingual evaluation.

3 Methodology

Section 3.1 presents our chosen datasets with a brief description of the structure, term frequency, and label distribution. We describe the general methodology, experimental setup, and the implementation details in Section 3.2 and 3.3. Finally, in Section 3.4 we describe the chosen evaluation metrics for the ATE task.

3.1 Dataset

The experiments were conducted on two datasets (ACTER [30] and RSDO5 version 1.1 [12]) containing texts from different languages and domains. The structures of both datasets are presented in Figure 1.

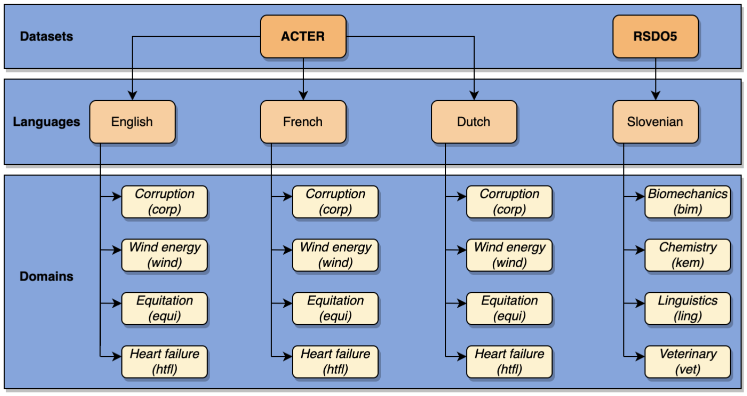


Fig. 1: The structure of RSDO5 and ACTER regarding languages and domains.

The ACTER dataset is a collection of 12 corpora covering four domains (Corruption (corp), Dressage (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 including both terms and named entities (NES), and the other one containing only terms (ANN). Table 1 summarizes the number of documents and unique terms for each domain. Note the discrepancy in size between the Heart failure domain and the other three domains, with the Heart failure domain containing the much more unique terms and documents[5].

Table 1: Number of documents and unique terms in the ACTER dataset.

Languages	Corruption (corp)			Equitation (equi)			Wind energy (wind)			Heart failure (htfl)		
	Docs	Terms		Docs	Terms		Docs	Terms		Docs	Terms	
		ANN	NES		ANN	NES		ANN	NES		ANN	NES
en	19	927	1,173	34	1,155	1,575	5	1,091	1,534	190	2,361	2,585
fr	19	979	1,207	78	961	1,181	2	773	968	210	2,228	2,374
nl	12	1,047	1,295	65	1,393	1,544	8	940	1,245	174	2,074	2,254

The second dataset is the RSDO5 corpus version 1.1 [12] containing texts in Slovenian (sl), a less-resourced Slavic language with rich morphology. Compiled during the course of the RSDO[6] national project, the RSDO5 corpus contains 12 documents (including three Ph.D. theses, a scientific book based on a Ph.D.

[5] The detailed description of ACTER can be found in the TermEval competition [30].
[6] https://www.cjvt.si/rsdo/en/project/

thesis, four graduate level textbooks, and four journal articles) with altogether about 250,000 words collected from diverse sources between 2000 to 2019 covering domains of Biomechanics (bim), Chemistry (kem), Veterinary (vet), and Linguistics (ling). The numbers of documents, tokens, and unique terms per domain are reported in Table 2. The documents from the Linguistics and Veterinary domains are longer (i.e., they have more tokens) and also contain more terms than Biomechanics and Chemistry. Most terms are made of one up to three words and only a few terms are longer than seven words. An example of a long term found in the corpus would be *"stojo po obračanju v nasprotni smeri urinega kazalca"* (stand after turning counterclockwise) in Biomechanics.

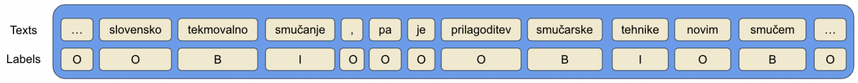
Table 2: Number of documents, tokens, and unique terms in the RSDO5 dataset.

Biomechanics (bim)			**Chemistry (kem)**			**Veterinary (vet)**			**Linguistics (ling)**		
Docs	**Tokens**	**Terms**	**Docs**	**Tokens**	**Terms**	**Docs**	**Tokens**	**Terms**	**Docs**	**Tokens**	**Terms**
3	61,344	2,319	3	65,012	2,409	3	75,182	4,748	3	109,050	4,601

Furthermore, both datasets contain several nested terms, i.e., a shorter term may appear within a larger term and vice versa. For example, in the RSDO5's Biomechanics domain, the term *"navor"* (torque) appears in terms such as *"sunek navora"* (torque shock), *"zunanji sunek navora"* (external torque shock), and *"izokinetični navor"* (isokinetic torque); in ACTER's English Corruption, term *"confiscation"* appears also in terms such as *"confiscation of corruption proceeds"*, *"confiscation of criminal assets"*, and *"confiscation of the proceeds of crime"*, 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.

3.2 Methodology

We experiment with XLM-RoBERTa, a Transformer-based model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. We consider ATE as a sequence-labeling task where the model returns a label for each token in a text sequence using the (B-I-O) labeling regime [32,19]. Here, 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 examples in Fig. 2).



Texts	...	slovensko	tekmovalno	smučanje	,	pa	je	prilagoditev	smučarske	tehnike	novim	smučem	...
Labels	O	O	B	I	O	O	O	O	B	I	O	B	O

Fig. 2: A sample of our labels in the RSDO5 corpus for term extraction.

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 evaluate the cross-domain performance of the model in a monolingual, cross-lingual, and multilingual setting. Altogether, 55 different scenarios 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. We fine-tune our model in a single language, which means we train three monolingual models for three languages (English, French, Dutch) and test each model in the same language as well as 12 monolingual models for Slovenian given 12 different combinations of train-validation-test split regarding the domains. This can be considered as a baseline to which we compare other settings.
2. **Cross-lingual setup.** We evaluate the capability of the model to apply the knowledge learned about ATE in one or more languages for ATE in another unseen language. Therefore, we fine-tune the ATE model in 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 all languages in the ACTER dataset (English, French, and Dutch) or using b.) training datasets from all 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. By doing so, we examine whether adding more data from other languages to the train set that matches 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 except the multilingual setting with additional Slovenian corpus in the training set where we use two domains from ACTER and all domains from RSDO5 corpus for the training. This way we want to check the generalization capabilities of the model, i.e. whether the knowledge the model obtained on one domain can be applied to the new, unseen domains, which would make the model applicable to arbitrary domains and therefore much more useful. In the ACTER dataset, we use Corruption and Wind energy domains as parts of the training, Equitation domains for validation, and Heart failure domain for testing in order to allow for a direct comparison with other benchmarks sharing the same train-validation-test setting [19], using the same dataset and evaluation setting (predicting on Heart Failure test set) from the related work. Meanwhile, in the RSDO5 corpus, we explore different train-validation-test combinations.

We divide the dataset into train-validation-test splits. The train split is used for fine-tuning the models while the validation split is used to prevent over-fitting during the fine-tuning phase. Finally, the test split is used for evaluation and is excluded during the model training. 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). 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 consider ATE as a sequence-labeling task and the models are trained to predict the labels from the (B-I-O) annotation scheme. The distribution across label types and the proportion of (B) and (I) labels in the total number of tokens per domain and per language are presented in Table 3. In the ACTER dataset, the proportion of terms in the texts is the largest for English, followed by French and then Dutch. The proportion of terms increases by from 1% upto 5% when adding NEs into the gold standards. In both datasets, the number of tokens annotated as terms (or parts of the term) only represents up to one-fourth (but usually much less) of the total tokens in the corpus, which means there is a significant imbalance between (B, I) tokens and tokens labeled as not terms (O).

Table 3: Label distribution and proportion of terms appearing per domain.

(a) The ACTER dataset.

Languages	**Corruption (corp)**				**Equitation (equi)**				**Wind energy (wind)**				**Heart failure (htfl)**			
	B	**I**	**O**	**% Term**	**B**	**I**	**O**	**% Term**	**B**	**I**	**O**	**% Term**	**B**	**I**	**O**	**% Term**
ANN en	4,558	2,200	44,287	13.24	10,745	1,938	46,215	21.53	5,046	3,323	49,873	14.37	9,819	4,504	41,522	25.65
ANN fr	4,461	2,823	51,918	12.30	8,420	2,373	50,487	17.61	5,928	4,405	43,976	19.03	7,165	4,027	43,976	20.29
ANN nl	4,251	1,517	46,730	10.99	10,243	1,509	45,011	20.70	4,174	826	50,642	8.99	8,529	1,391	45,142	18.02
NES en	6,050	3,226	41,769	18.17	11,340	2,377	45,181	23.29	6,040	4,111	48,091	17.43	10,115	4,855	40,875	26.81
NES fr	6,021	3,996	49,185	16.92	8,699	2,632	49,949	18.49	7,356	4,524	53,868	18.07	7,394	4,172	43,602	20.97
NES nl	5,585	2,308	44,605	15.03	10,416	1,625	44,722	21.21	4,708	1,084	49,850	10.41	8,770	1,627	44,665	18.88

(b) The 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**
sl	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

We employ the XLM-RoBERTa token classification model and its "fast" XLM-RoBERTa tokenizer from the Huggingface library[7]. We fine-tune the model for up to 20 epochs (i.e., we employ the early stopping regime) 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. The documents are first split into sentences. Then, 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).

[7] https://huggingface.co/models

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.

3.4 Evaluation Metrics

We evaluate the performance of the ATE system by comparing the candidate list extracted on the whole test set level with the manually annotated gold standard of each domain using strictly matching with Precision (P), Recall (R), and F1-score (F1). These evaluation metrics have also been used in the related work, including the TermEval 2020 [11,30,19] and Slovenian benchmark [24]. Therefore, our results are directly comparable to the SOTA methods.

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 SOTAs from the related work.

4.1 Prediction on the ACTER test set

Table 4 demonstrates the performance of XLM-RoBERTa on the cross-domain sequence-labeling ATE task on the ACTER test set in the monolingual, cross-lingual, and multilingual setting. We group the results according to the test language in the ACTER corpora for better comparison among settings. The results indicate that cross- and multilingual models surpass the performance of the monolingual ones according to all evaluation metrics except for when it comes to the Precision obtained by the French monolingual model on the French test set. Multilingual models tend to outperform cross-lingual ones, except for the cross-lingual model trained in Dutch and applied to the English test set. This multilingual model boosts the F1-score performance by up to 2% in ANN and 1% in the NES task when compared to the second-highest-performing model. 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, which, on average, increases by 18.17% in ANN and 13.54% in NES test set compared with the monolingual setting.

Table 5 presents a comparison between the best-performing models in this work in terms of F1-score and the benchmark approaches in the ACTER dataset, including the solutions from the winning teams in the competition (TALN-LS2N [11] won on the English and French test set while NLPLab UQAM [20] won on the Dutch test set) and other methods proposed in Rigouts et al. [32] and Lang et al. [19], which are described in Section 2. Note that all the approaches from the related work are cross-domain and use the Heart failure domain as the test set and the rest of the data for training or validation. For the ANN task in

Table 4: Evaluation on ACTER given Heart failure as test set.

Train language	Enlish test set						French test set						Dutch test set					
	ANN			NES			ANN			NES			ANN			NES		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
en	58.08	48.12	52.63	62.07	52.03	56.61	66.69	47.89	55.75	70.63	53.79	**61.07**	69.23	61.09	64.91	72.95	63.04	67.63
fr	56.94	33.21	41.95	60.01	39.07	47.33	**70.51**	44.43	54.51	**72.41**	48.53	58.11	72.12	51.01	59.76	73.63	55.50	63.29
nl	55.64	56.37	**56.00**	57.60	58.34	**57.97**	66.49	51.48	58.03	67.60	53.16	59.52	70.25	62.15	65.95	73.29	61.49	66.87
en, fr	57.16	51.21	54.02	60.43	51.45	55.58	63.70	52.38	57.49	68.13	52.78	59.48	72.52	61.72	66.69	73.08	63.49	67.95
en, nl	58.00	48.67	52.93	**62.39**	51.33	56.32	65.25	44.17	52.68	68.67	52.36	59.42	69.29	60.17	64.41	74.35	61.71	67.44
fr, nl	**60.84**	46.84	52.93	62.27	50.37	55.69	69.20	48.29	56.88	70.72	49.54	58.26	**75.72**	56.70	64.84	**76.74**	59.58	67.08
en, fr, nl	56.83	53.03	54.86	60.76	52.53	56.35	68.01	50.67	58.07	48.30	65.57	55.63	69.92	64.32	67.00	73.66	62.91	67.86
en, fr, nl, sl	45.88	**66.29**	54.23	48.30	**65.57**	55.63	58.10	**61.62**	**59.81**	59.48	**62.51**	60.96	62.74	**75.51**	**68.54**	63.57	**73.69**	**68.26**

English and Dutch and the NES task in French, our methods outperform other approaches in terms of F1-score. Despite not surpassing the SOTA in the French ANN task and the other two NES tasks, our method still offers competitive performance being outperformed by the HAMLET approach [32] with a small margin of 0.39% in ANN French, and by the token classifier [19] with about 0.33% in NES English. In terms of multilingual evaluation, we show that in contrast to the findings of Lang et al. [19], adding different languages in general slightly improves the models.

Table 5: F1-score comparison between our results and related work in ACTER.

Methods	English		French		Dutch	
	ANN	NES	ANN	NES	ANN	NES
Winning teams [11]	44.99	46.66	45.94	48.15	18.60	18.70
HAMLET [32]	54.20	55.40	**60.20**	60.80	66.10	66.00
Sequence Classifier [19]	x	46.00	x	48.10	x	58.00
NMT [19]	x	55.30	x	57.60	x	59.60
Token classifer [19]	x	**58.30**	x	57.60	x	**69.80**
NMF-based approaches [26]	33.50	33.70	30.90	30.70	30.10	30.30
Our best classifers	**56.00**	57.97	59.81	**61.07**	**68.54**	68.26

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 are presented in Table 6. 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, 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 3.1 might be one of the factors that contribute to this behavior. Moreover, a significant performance boost can be observed for the Linguistics domain when the model is trained in the Chemistry 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%.

Table 6: The evaluation of monolingual and multilingual learning in RSDO5.

Validation	Testing	Monolingual setup			multilingual setup					
					ANN			NES		
		P	R	F1	P	R	F1	P	R	F1
vet	ling	**69.55**	64.05	66.69	67.68	69.55	68.60	67.19	**69.88**	**68.51**
bim	ling	69.48	**73.66**	**71.51**	**69.78**	66.16	67.92	67.81	68.53	68.17
kem	ling	66.20	72.38	69.15	66.50	**71.35**	**68.84**	**67.89**	69.03	68.46
ling	vet	71.06	66.72	68.82	**70.96**	65.27	68.00	69.22	67.40	68.30
kem	vet	**72.66**	65.59	**68.94**	69.75	**68.83**	**69.29**	**70.49**	**67.75**	**69.09**
bim	vet	69.30	**68.07**	68.68	69.77	68.43	69.09	69.26	64.72	66.91
ling	kem	68.67	55.13	61.16	68.26	59.28	63.45	67.54	54.59	60.38
bim	kem	70.14	**60.27**	**64.83**	69.63	**61.19**	**65.14**	**69.25**	52.72	59.86
vet	kem	**70.23**	59.24	64.27	**69.90**	58.41	63.64	67.92	**59.24**	**63.28**
vet	bim	**63.51**	**66.80**	**65.11**	61.14	**64.94**	**62.98**	60.94	66.67	63.68
ling	bim	62.25	65.20	63.69	60.53	63.82	62.13	**62.62**	62.27	62.44
kem	bim	62.35	63.99	63.16	**65.71**	59.16	62.26	61.78	**67.05**	**64.31**

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 6 demonstrates the comparative performance of the multilingual and the monolingual approaches, which is consistent with the results in the prediction of the ACTER test set.

Table 7: 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
Monolingual	**69.48**	**73.66**	**71.51**	**72.66**	65.59	68.94	**70.14**	60.27	64.83	**63.51**	66.80	**65.11**
Multilingual	66.50	71.35	68.84	69.75	**68.83**	**69.29**	69.63	**61.19**	**65.14**	61.78	**67.05**	64.31
SOTA [24]	52.20	25.40	34.10	66.90	19.30	29.90	47.80	31.40	37.80	53.80	24.80	33.90

Furthermore, in Table 7, we present the results from the related work for the RSDO5 dataset [24] in comparison to the proposed monolingual and multilingual approaches. The results from [24]'s method are taken from Hanh et al. [33]. In general, our approach outperforms the approach proposed in Ljubevsic et al. [24] by a large margin on all domains and according to all evaluation metrics, especially when it comes to Recall. Overall, we achieve results roughly twice as high as the approach proposed by Ljubevsic et al. [24] in terms of F1-score for all test domains regarding both monolingual and multilingual learning. We show that the multilingual experiments do in several cases improve our monolingual results [33], but this is not systematic.

5 Error analysis

In order to determine whether the term length affects the models' performance, we calculate Precision and Recall separately for terms of length k = {1,2,3,4,

≥5}. The number of predicted candidate terms (Preds), ground truth (GT), correct predictions (TPs), Precision, and Recall regarding different term lengths k and test domains are presented in Table 8. The results for ACTER's dataset (Table 8a) were obtained by employing the best performing model for a specific language in terms of F1-score on the Heart failure test set. The results for the RSDO5 dataset (Table 8b) were obtained by employing the best-performing model for a specific test domain in F1-score.

Table 8: Performance per term length per domain in each test set.

(a) ACTER test set.

k	Enlish					French					Dutch				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	1,009	1,170	639	63.33	54.62	1,153	1,309	829	71.90	63.33	2,005	1,687	1,292	64.44	76.59
2	985	801	501	50.86	62.55	490	620	320	65.31	51.61	661	391	303	45.84	77.49
3	553	377	256	46.29	67.90	163	266	100	61.35	37.59	108	108	55	50.93	50.93
4	163	142	86	52.76	60.56	47	91	24	51.06	26.37	19	35	10	52.63	28.57
≥5	53	95	26	49.06	27.37	13	88	4	30.77	4.55	1	33	1	100.00	3.03

(b) RSDO5 Linguistics test set.

k	Linguistics					Veterinary					Chemistry					Biomechanics				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	2,078	1,728	1,300	62.56	75.23	2,159	2,067	1,472	68.18	71.21	943	890	580	61.51	65.17	1,079	718	22	48.38	72.70
2	2,631	2,404	1,858	70.62	77.29	2,062	2,103	1,448	70.22	68.85	1,073	1,202	768	71.58	63.89	1,153	1,172	822	71.29	70.14
3	322	360	7,191	59.32	53.06	314	446	182	57.96	40.81	164	260	93	56.71	35.77	223	286	124	55.61	43.36
4	57	80	31	54.39	38.75	28	77	10	35.71	12.99	26	46	11	42.31	23.91	26	59	11	42.31	18.64
≥5	12	29	79	75.00	31.03	3	55	2	66.67	3.64	3	11	0	0.00	0.00	11	84	5	45.45	5.95

The models proved to be good at predicting terms containing up to four words for English and Dutch and up to three words for French. The results on the RSDO5 dataset are similar, showing that the models are good at predicting short terms containing up to three words for all four domains of the RSDO5 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 75.00% Precision and a decent 31.03% 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 it comes to predictions in the Chemistry domain, there are no correct term predictions that consist of more than five words.

6 Conclusion

In summary, we investigated the possibilities of cross- and multilingual learning compared to the monolingual setting in the cross-domain sequence-labeling term extraction given the experiments conducted on multi-domain corpora, namely the ACTER and RSDO5 datasets. We also evaluated the 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. In addition, we examined the cross-lingual effect of rich-resourced training language on less-resourced testing one such as Slovenian. The results demonstrate a promising impact of multilingual and cross-lingual cross-domain learning that outperforms the related works in both datasets, which proves their potential when transferring from the rich- to the less-resourced languages.

However, we believe that there remains room for improvement in the field of supervised term extraction. In the future, 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.

7 Acknowledgements

The work was partially supported by the Slovenian Research Agency (ARRS) core research programme Knowledge Technologies (P2-0103), as well as the Ministry of Culture of Republic of Slovenia through project Development of Slovene in Digital Environment (RSDO). The first author was partly funded by Region Nouvelle Acquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: 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 (2019)
2. Amjadian, E., Inkpen, D., Paribakht, T., Faez, F.: Local-Global Vectors to Improve Unigram Terminology Extraction. In: Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016). pp. 2–11 (2016)
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: ACL (2020)
4. Daille, B., Gaussier, É., Langé, J.M.: Towards Automatic Extraction of Monolingual and Bilingual Terminology. In: COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics (1994)
5. Damerau, F.J.: Evaluating computer-generated domain-oriented vocabularies. Information processing & management **26**(6), 791–801 (1990)
6. ElKishky, A., Song, Y., Wangx, C., Voss, C.R., Han, J.: Scalable topical phrase mining from text corpora. Proceedings of the VLDB Endowment **8**(3), 305–316 (2014)
7. Erjavec, T., Fišer, D., Ljubešić, N.: The kas corpus of slovenian academic writing. Language Resources and Evaluation **55**(2), 551–583 (2021)

8. Fišer, D., Suchomel, V., Jakubícek, M.: Terminology extraction for academic slovene using sketch engine. In: Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016. pp. 135–141 (2016)
9. Frantzi, K.T., Ananiadou, S., Tsujii, J.: 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 (1998)
10. Gao, Y., Yuan, Y.: Feature-less End-to-end Nested Term extraction. In: CCF International Conference on Natural Language Processing and Chinese Computing. pp. 607–616. Springer (2019)
11. Hazem, A., Bouhandi, M., Boudin, F., Daille, B.: TermEval 2020: TALN-LS2N System for Automatic Term Extraction. In: Proceedings of the 6th International Workshop on Computational Terminology. pp. 95–100 (2020)
12. Jemec Tomazin, M., Trojar, M., Atelšek, S., Fajfar, T., Erjavec, T., Žagar Karer, M.: Corpus of term-annotated texts RSDO5 1.1 (2021), `http://hdl.handle.net/11356/1470`, slovenian language resource repository CLARIN.SI
13. Jemec Tomazin, M., Trojar, M., Žagar, M., Atelšek, S., Fajfar, T., Erjavec, T.: Corpus of term-annotated texts rsdo5 1.0 (2021)
14. Justeson, J.S., Katz, S.M.: Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. Natural language engineering **1**(1), 9–27 (1995)
15. Kessler, R., Béchet, N., Berio, G.: Extraction of terminology in the field of construction. In: 2019 First International Conference on Digital Data Processing (DDP). pp. 22–26. IEEE (2019)
16. Koloski, B., Pollak, S., Škrlj, B., Martinc, M.: Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? arXiv preprint arXiv:2202.06650 (2022)
17. Kucza, M., Niehues, J., Zenkel, T., Waibel, A., Stüker, S.: Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In: INTERSPEECH. pp. 2072–2076 (2018)
18. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: 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 (2016)
19. Lang, C., Wachowiak, L., Heinisch, B., Gromann, D.: 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 (2021)
20. Le, N.T., Sadat, F.: Multilingual automatic term extraction in low-resource domains. In: The International FLAIRS Conference Proceedings. vol. 34 (2021)
21. Le Serrec, A., L'Homme, M.C., Drouin, P., Kraif, O.: 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 (2010)
22. Lingpeng, Y., Donghong, J., Guodong, Z., Yu, N.: Improving retrieval effectiveness by using key terms in top retrieved documents. In: European Conference on Information Retrieval. pp. 169–184. Springer (2005)
23. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)

24. Ljubešić, N., Fišer, D., Erjavec, T.: Kas-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. In: International Conference on Text, Speech, and Dialogue. pp. 115–126. Springer (2019)
25. Martinc, M., Škrlj, B., Pollak, S.: Tnt-kid: Transformer-based neural tagger for keyword identification. Natural Language Engineering p. 1–40 (2021). https://doi.org/10.1017/S1351324921000127
26. Nugumanova, A., Akhmed-Zaki, D., Mansurova, M., Baiburin, Y., Maulit, A.: Nmf-based approach to automatic term extraction. Expert Systems with Applications **199**, 117179 (2022)
27. Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostaja, T., Vintar, Š., Fišer, D.: Extracting data from comparable corpora. In: Using Comparable Corpora for Under-Resourced Areas of Machine Translation, pp. 89–139. Springer (2019)
28. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)
29. Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., Pollak, S.: 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 (2019)
30. Rigouts Terryn, A., Hoste, V., Drouin, P., Lefever, E.: 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) (2020)
31. Rigouts Terryn, A., Hoste, V., Lefever, E.: In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and Evaluation **54**(2), 385–418 (2020)
32. Rigouts Terryn, A., Hoste, V., Lefever, E.: HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. Terminology (2021)
33. Tran, H.T.H., Martinc, M., Doucet, A., Pollak, S.: A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction. In: Submitteed to Slovenian conference on Language Technologies and Digital Humanities (2022, under review)
34. Tran Hanh, T.H., Doucet, A., Sidere, N., Moreno, J.G., Pollak, S.: 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 (2021)
35. Vintar, S.: 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 (2010)
36. Wolf, P., Bernardi, U., Federmann, C., Hunsicker, S.: From statistical term extraction to hybrid machine translation. In: Proceedings of the 15th Annual conference of the European Association for Machine Translation (2011)
37. Zhang, Z., Gao, J., Ciravegna, F.: 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 (2018)