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

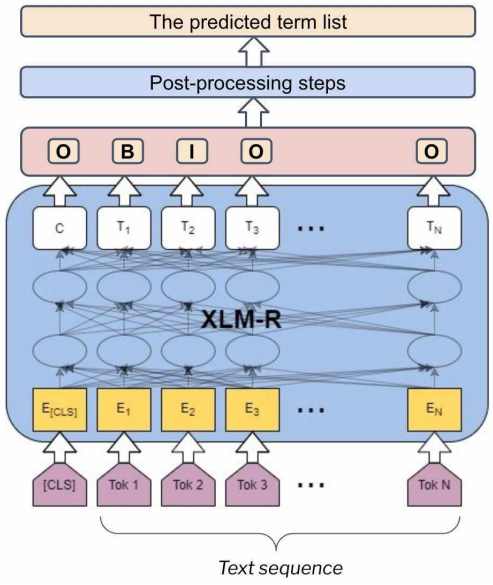

Figure 2: The overall architecture.

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

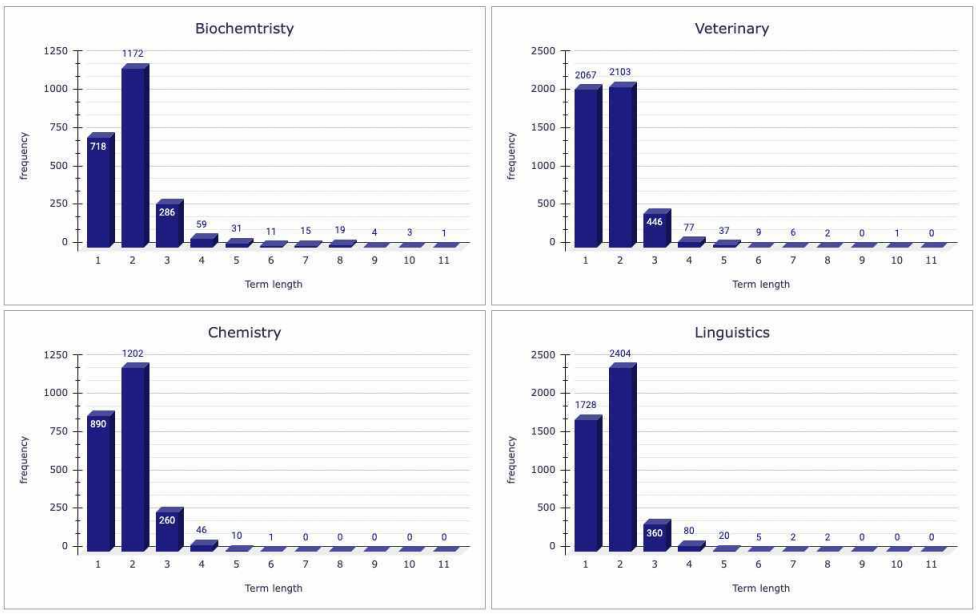

Figure 3: The frequencies of terms of specific length per each domain in a Slovenian dataset.

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*.