








Ensembling Transformers for Cross-domain Automatic Term Extraction

Hanh Thi Hong Tran^{1,2,3} , Matej Martinc¹ , Andraz Pelicon¹ ,
Antoine Doucet³ , and Senja Pollak² 

¹ Jožef Stefan International Postgraduate School,
Jamova Cesta 39, 1000 Ljubljana, Slovenia
tran.hanh@ijs.si

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 plays an essential role in domain language understanding and several natural language processing downstream tasks. In this paper, we propose a comparative study on the predictive power of Transformers-based pretrained language models toward term extraction in a multi-language cross-domain setting. Besides evaluating the ability of monolingual models to extract single- and multi-word terms, we also experiment with ensembles of mono- and multilingual models by conducting the intersection or union on the term output sets of different language models. Our experiments have been conducted on the ACTER corpus covering four specialized 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). The results show that the strategy of employing monolingual models outperforms the state-of-the-art approaches from the related work leveraging multilingual models, regarding all the languages except Dutch and French if the term extraction task excludes the extraction of named entity terms. Furthermore, by combining the outputs of the two best performing models, we achieve significant improvements.

Keywords: Automatic term extraction · ATE · Low resource · ACTER · RSDO5 · Monolingual · Cross-domain

1 Introduction

Automatic Term Extraction (ATE) is the task of identifying specialized terminology from the domain-specific corpora. By easing the time and effort needed to manually extract the terms, ATE is not only widely used for terminographical tasks (e.g., glossary construction [26], specialized dictionary creation [22], etc.) but it also contributes to several complex downstream tasks (e.g., machine translation [40], information retrieval [23], sentiment analysis [28], to cite a few).

With recent advances in natural language processing (NLP), a new family of deep neural approaches, namely Transformers [38], has been pushing the state-of-the-art (SOTA) in several sequence-labeling semantic tasks, e.g., named entity recognition (NER) [18, 37] and machine translation [41], among others. The TermEval 2020 Shared Task on Automatic Term Extraction, organized as part of the CompuTerm workshop [31], presented one of the first opportunities to systematically study and compare various ATE systems with the advent of The Annotated Corpora for Term Extraction Research (ACTER) dataset [31, 32], a novel corpora covering four domains and three languages. Regarding Slovenian, the RSDO5¹ corpus [13] was created with texts from four specialized domains. Inspired by the success of Transformers for ATE in the TermEval 2020, we propose an extensive study of their performance in a cross-domain sequence-labeling setting and evaluate different factors that influence extraction effectiveness. The experiments are conducted on two datasets: ACTER and RSDO5 corpora.

Our major contributions can be summarized as the three following points:

- An empirical evaluation of several monolingual and multilingual Transformer-based language models, including both masked (e.g., BERT and its variants) and autoregressive (e.g., XLNet) models, on the cross-domain ATE tasks;
- Filling the research gap in ATE task for Slovenian by experimenting with different models to achieve a new SOTA in the RSDO5 corpus.
- An ensembling Transformer-based model for ATE that further improves the SOTA in the field.

This paper is organised as follows: Sect. 2 presents the related work in term extraction. Next, we introduce our methodology in Sect. 3, including the dataset description, the workflow and experimental settings, as well as the evaluation metrics. The corresponding results are presented in Sect. 4. Finally, we conclude the paper and present future directions in Sect. 5.

2 Related Work

The research into monolingual ATE was first introduced during the 1990s [6, 15] and the methods at the time included the following two-step procedure: (1) extracting a list of candidate terms; and (2) determining which of these candidate terms are correct using either supervised or unsupervised techniques. We briefly summarize different supervised ATE techniques according to their evolution below.

2.1 Approaches Based on Term Characteristics and Statistics

The first ATE approaches leveraged linguistic knowledge and distinctive linguistic aspects of terms to extract a possible candidate list. Several NLP techniques are employed to obtain the term’s linguistic profile (e.g., tokenization, lemmatization, stemming, chunking, etc.). On the other hand, several studies proposed

¹ <https://www.clarin.si/repository/xmlui/handle/11356/1470>.

statistical approaches toward ATE, mostly relying on the assumption that a higher candidate term frequency in a domain-specific corpus (compared to the frequency in the general corpus) implies a higher likelihood that a candidate is an actual term. Some popular statistical measures include termhood [39], unit-hood [5] or C-value [10]. Many current systems still apply their variations or rely on a hybrid approach combining linguistic and statistical information [16,30].

2.2 Approaches Based on Machine Learning and Deep Learning

The recent advances in word embeddings and deep neural networks have also influenced the field of term extraction. Several embeddings have been investigated for the task at hand, e.g., non-contextual [1,43], contextual [17] word embeddings, and the combination of both [11]. The use of language models for ATE tasks is first documented in the TermEval 2020 [31] on the trilingual ACTER dataset. While the Dutch corpus winner used BiLSTM-based neural architecture with GloVe word embeddings, the English corpus winner [12] fed all possible extracted n-gram combinations into a BERT binary classifier. Several Transformer variations have also been investigated [12] (e.g., BERT, RoBERTa, CamemBERT, etc.) but no systematic comparison of their performance has been conducted. Later, the HAMLET approach [33] proposed a hybrid adaptable machine learning system that combines linguistic and statistical clues to detect terms. Recently, sequence-labeling approaches became the most popular modeling option. They were first introduced by [17] and then employed by [20] to compare several ATE methods (e.g., binary sequence classifier, sequence classifier, token classifier). Finally, cross-lingual sequence labeling proposed in [4,20,35] demonstrates the capability of multilingual models and the potential of cross-lingual learning.

2.3 Approaches for Slovenian Term Extraction

The ATE research for the less-resourced languages, especially Slovenian, is still hindered by the lack of gold standard corpora and the limited use of neural methods. Regarding the corpora, the recently compiled Slovenian KAS corpus [8] was quickly followed by the domain-specific RSDO5 corpus [14]. Regarding the methodologies, techniques evolved from purely statistical [39] to more machine learning based approaches. For example, [25] extracted the initial candidate terms using the CollTerm tool [29], a rule-based system employing a language-specific set of term patterns from the Slovenian SketchEngine module [9]. The derived candidate list was then filtered using a machine learning classifier with features representing statistical measures. Another recent approach [30] focused on the evolutionary algorithm for term extraction and alignment. Finally, [36] was one of the first to explore the deep neural approaches for Slovenian term extraction, employing XLMRoBERTa in cross- and multilingual settings.

3 Methods

We briefly describe our chosen datasets in Sect. 3.1, the general methodology in Sect. 3.2 and the chosen evaluation metrics in Sect. 3.3.

3.1 Datasets

The experiments have been conducted on two datasets: ACTER v1.5 [31] and RSDO5 v1.1 [13]. The ACTER dataset is a manually annotated 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). It has two versions of gold standard annotations: one including both terms and named entities (NES), and the other containing only terms (ANN). Meanwhile, the RSDO5 corpus v1.1 [13] includes texts in Slovenian (sl), a less-resourced Slavic language with rich morphology. Compiled during the RSDO national project, the corpus contains 12 documents covering four domains, Biomechanics (bim), Chemistry (kem), Veterinary (vet), and Linguistics (ling).

3.2 Workflow

We consider ATE as a sequence-labeling task [35] with IOB labeling regime [20, 33]. The model is first trained to predict a label for each token in the input text sequence, and then applied to the unseen test data. From the token sequences labeled as terms, the final candidate term list for the test data is composed.

3.2.1 Empirical Evaluation of Pretrained Language Models

We conduct a systematic evaluation of mono- and multilingual Transformers-based models on the ATE task modeled as sequence labeling. The models were obtained from Huggingface² according to the number of downloads and likes criteria. The chosen models are presented in Fig. 1. Regarding the multilingual systems, we investigate the performance of mBERT [7] (*bert-base-multilingual-uncased*), mDistilBERT [34] (*distilbert-base-multilingual-cased*), InforXLM [2] (*microsoft/ infoxlm-base*), and XLMRoBERTa [3] (*xlm-roberta-base*). All the chosen multilingual models are fine-tuned in a monolingual fashion due to findings from the related work [20, 35] showing that no (or only marginal) gains are obtained if the model is fine-tuned on the multilingual training data.

Regarding the monolingual models, we evaluate several English autoencoding Transformer-based models, including ALBERT [19] (*albert-base-v1* and *albert-base-v2*), BERT [7] (*bert-base-uncased*), DistilBERT [34] (*distilbert-base-uncased*), ELECTRA (*electra-small-generator*) and RoBERTa [24] (*xlm-roberta-base*), and one autoregressive model, XLNet [42] (*xlnet-base-cased*). For French, we use CamemBERT [27] (*camembert-base*) and FlauBERT [21] (*flaubert_base_uncased*), for Dutch, we employ BERTje (*bert-base-dutch-cased*) and RobBERT (*robBERT-base* and *robbert-v2-dutch-base*) models, and for Slovenian, we choose SloBERTa (*sloberta*), the RoBERTa-based model trained on a large Slovenian corpus.

² <https://huggingface.co/models>.

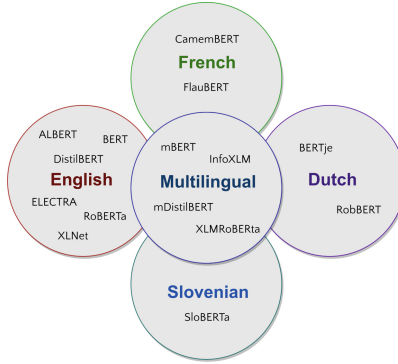


Fig. 1. Empirical evaluation of pre-trained language models on the ATE task.

3.2.2 Ensemble of Transformer Models

Regarding results in Sect. 3.2.1, we propose a novel ensembling approach based on Transformer models for ATE task as we observe the general tendency for Precision to be better than Recall for all but few monolingual and multilingual models tested (see Tables 1 and 2). This leads us to believe that by combing the outputs of different models, we could achieve improvements in Recall and by extension also in the overall F1-score. We consider two strategies for combining the outputs from different models of the ensemble, namely the union and the intersection of the candidate term lists from the models of the ensemble. See the entire procedure in Fig. 2.

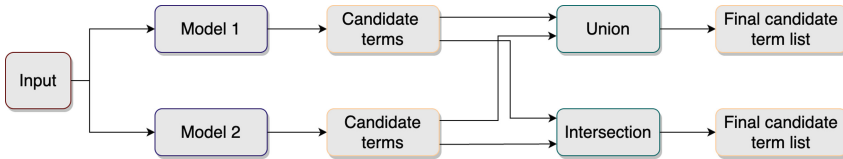


Fig. 2. The general ensembling workflow.

We hypothesize that by combining the outputs of two models, we might be able to significantly improve the Recall of the term extraction system. To validate this hypothesis, we test three combinations: Combine the outputs of the (1) best mono- and multilingual models; (2) two best monolingual models; and (3) two best multilingual models.

3.3 Evaluation Metrics

We evaluate each term extraction system by comparing the aggregated list of candidate terms extracted on the level of the whole test set with the manually annotated gold standard term list using Precision, Recall, and F1-score. These evaluation metrics have also been used in the related work [12, 20, 31].

4 Results

We first present the results of mono- and multilingual Transformer-based models obtained on ACTER and RSDO5 test sets compared with the SOTAs. Then, we demonstrate the impact of the ensemble post-processing step.

4.1 Monolingual Evaluation

4.1.1 ACTER Corpus

Not many approaches have been tested on the ACTER corpus v1.5 due to its novelty. Thus, we apply the approach proposed by [20] (i.e., employing XLM-RoBERTa as a token classifier), which achieved SOTA on the previous corpus version, and consider it as a baseline. The Heart failure domain is used as a test set, same as in TermEval 2020.

Table 1. Results of monolingual term extraction on the ACTER dataset.

	Models	ANN			NES		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Mono	albert-base-v1	52.58	47.40	49.86	54.42	54.63	54.52
	albert-base-v2	49.85	48.50	49.17	57.01	55.13	56.05
	bert-base-uncased	59.06	32.44	41.88	61.42	47.50	53.57
	distilbert-base-uncased	58.24	38.75	46.54	61.06	48.24	53.90
	electra-small-generator	56.46	46.80	51.18	58.17	47.31	52.18
	roberta-base	58.10	51.04	54.34	62.28	56.30	59.14
	xlnet-base-cased	56.50	53.92	55.18	58.34	57.30	57.82
Multi	bert-base-multilingual-uncased	55.21	35.24	43.02	62.06	49.44	55.04
	distilbert-base-multilingual-cased	55.14	45.45	49.83	57.10	54.20	55.61
	infoclm-base	57.67	54.64	56.11	61.18	54.48	57.64
	xlm-roberta-base (baseline)	57.34	51.46	54.24	58.80	55.52	57.11
(a) English corpus							
	Models	ANN			NES		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Mono	camembert-base	70.51	44.97	54.92	70.74	52.23	60.09
	flauberta	75.91	26.17	38.92	75.28	39.01	51.39
Multi	bert-base-multilingual-uncased	67.77	37.66	48.42	69.39	48.99	57.43
	distilbert-base-multilingual-cased	64.45	43.45	51.91	65.20	48.78	55.81
	infoclm-base	68.74	39.77	50.39	71.10	48.90	57.95
	xlm-roberta-base (baseline)	68.85	48.61	56.99	70.71	46.46	56.08
(b) French corpus							
	Models	ANN			NES		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Mono	bert-base-dutch-cased	65.59	65.53	65.56	67.61	66.02	66.81
	roBERT-base	69.58	36.84	48.17	71.63	55.01	62.23
	robert-v2-dutch-base	71.56	36.40	48.25	73.58	55.72	63.42
Multi	bert-base-multilingual-uncased	70.67	62.49	66.33	72.34	63.71	67.75
	distilbert-base-multilingual-cased	69.80	61.28	65.26	69.45	66.15	67.76
	infoclm-base	70.43	66.73	68.53	73.47	64.24	68.55
	xlm-roberta-base (baseline)	68.53	67.94	68.23	73.93	60.65	66.63
(c) Dutch corpus							

In general, multilingual pretrained models outperform the monolingual ones in Recall and F1-score when applied for extraction of the ANN annotations in all three languages. If named entities are included (NES), monolingual models outperform multilingual models in two (English and French) out of three languages in the ACTER dataset. When it comes to individual models, InfoXLM

outperforms other mono- and multilingual models in the F1-score on the Dutch corpus (for both ANN and NES) and on the English corpus (for ANN). If we compare the results of our study with the XLMRoBERTa baseline using the same monolingual settings from [20], our best-performing models surpass the baseline in all cases (e.g., the F1-score increases by 1.87% on ANN and 1.5% on NES in the English corpus; 4.01% on French NES; 0.3% on ANN and 1.92% on NES in the Dutch corpus) except for the French ANN annotations.

4.1.2 RSDO5 Corpus

We also compare the performance of different mono- and multilingual models on the RSDO5 corpus. Here, we evaluate the models on all domains as demonstrated in Table 2. By using two domains from the RSDO5 corpus for training, the third one for validation, and the last one for testing, all the models prove to have relatively consistent performance across different combinations. The monolingual SloBERTa model outperforms other approaches (including the XLMRoBERTa baseline from [36]) in all cases by a relatively large margin in F1-score. By employing this model and looking at the best performing train/validation combinations for each test domain, we improve the SOTA baseline in the Linguistics domain by 2.21%, in Veterinary by 2.35%, in Chemistry by 5.26%, and in Biomechanics by 2.66% regarding F1-score. Our results, thus, set a new SOTA on the Slovenian corpus.

Table 2. Results of monolingual term extraction on the RSDO5 dataset.

Training	Val	Test	xlm-roberta-base			sloberta			infxlm-base		
			Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
bim + kem	vet	ling	69.55	64.05	66.69	73.23	70.51	71.84	68.37	71.38	69.84
bim + vet	kem	ling	66.20	72.38	69.15	73.91	73.53	73.72	67.74	71.46	69.55
kem + vet	bim	ling	69.48	73.66	71.51	74.45	73.96	74.20	73.71	66.90	70.14
bim + kem	ling	vet	71.06	66.72	68.82	77.56	65.96	71.29	71.04	63.69	67.16
bim + ling	kem	vet	72.66	65.59	68.94	78.33	65.31	71.23	66.88	68.93	67.89
ling + kem	bim	vet	69.30	68.07	68.68	76.66	64.89	70.29	72.69	63.63	67.86
bim + vet	ling	kem	68.67	55.13	61.16	72.14	65.88	68.87	67.77	60.40	63.87
bim + ling	vet	kem	70.23	59.24	64.27	70.29	68.45	69.36	72.00	56.58	63.37
ling + vet	bim	kem	70.14	60.27	64.83	73.52	66.96	70.09	71.22	59.49	64.83
vet + kem	ling	bim	62.25	65.20	63.69	67.97	67.36	67.66	63.60	60.59	62.06
vet + ling	kem	bim	62.35	63.99	63.16	68.97	66.62	67.77	56.66	67.53	61.62
ling + kem	vet	bim	63.51	66.80	65.11	67.15	67.79	67.47	60.61	64.04	62.28

Training	Val	Test	bert-base-multilingual-uncased			distilbert-base-multilingual-cased		
			Precision	Recall	F1-score	Precision	Recall	F1-score
bim + kem	vet	ling	66.77	65.86	66.31	61.82	53.38	57.29
bim + vet	kem	ling	66.80	68.01	67.40	59.14	67.20	62.91
kem + vet	bim	ling	65.97	69.62	67.75	60.94	58.16	59.52
bim + kem	ling	vet	68.18	61.56	64.70	63.76	58.70	61.13
bim + ling	kem	vet	68.58	65.46	66.98	65.83	58.15	61.75
ling + kem	bim	vet	69.12	60.61	64.59	66.01	54.02	59.42
bim + vet	ling	kem	65.35	59.73	62.41	55.73	60.52	58.03
bim + ling	vet	kem	65.53	63.22	64.35	60.15	55.83	57.91
ling + vet	bim	kem	67.32	53.96	59.90	59.53	57.70	58.60
vet + kem	ling	bim	62.63	60.85	61.73	57.84	55.84	56.82
vet + ling	kem	bim	65.25	58.30	61.58	60.62	56.36	58.41
ling + kem	vet	bim	62.69	63.61	63.15	62.04	52.44	56.84

4.2 Transformer Ensembling

We also evaluate the performance of the proposed ensembling approach described in Sect. 3.2.2. The improvements/decline in performance over the best single model on different languages of the ACTER dataset are shown in Fig. 3. The results indicate that combining the acquired term sets of the two best-performing classifiers (no matter what type of classifiers they are) using the union always results in the biggest gain.

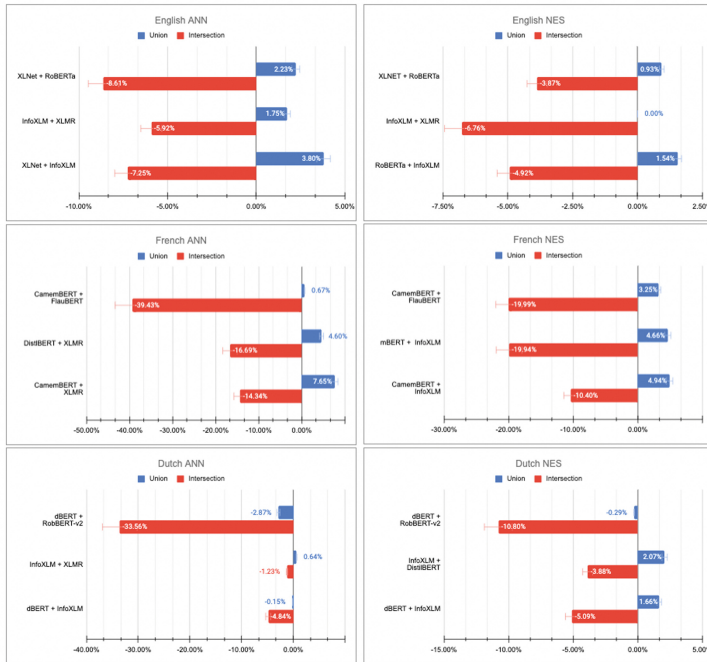


Fig. 3. F1-score improvement by combining two best classifiers in ACTER.

5 Conclusion

We proposed an empirical evaluation of different mono- and multilingual Transformers based models on the monolingual sequence-labeling cross-domain term extraction. The experiments were conducted on the trilingual ACTER dataset and the Slovenian RSDO5 dataset. Furthermore, we tested how ensembling different mono- or multilingual models affects the performance of the overall term extractor. The results demonstrate that multilingual models outperform the monolingual ones in Recall and F1-score when applied for ANN extraction. Meanwhile, monolingual models capture the information about terms better than multilingual ones when it comes to the extraction of NES annotations. We also

showed that by ensembling different Transformer models we can obtain further boosts in performance for all languages. As a consequence, we established the new SOTA on the ACTER and RSDO5 datasets.

In the future, we would like to take advantage of prompt engineering by considering ATE as a language model ranking problem in a sequence-to-sequence framework, where original sentences and statement templates filled by candidate terms are regarded as the source sequence and the target.

Acknowledgements. The work was partially supported by the Slovenian Research Agency (ARRS) core research programme Knowledge Technologies (P2-0103), and 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.

References

1. 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)
2. Chi, Z., et al.: InfoXLM: an information-theoretic framework for cross-lingual language model pre-training. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3576–3588 (2021)
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint [arXiv:1911.02116](https://arxiv.org/abs/1911.02116) (2019)
4. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: ACL (2020)
5. 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)
6. Damerau, F.J.: Evaluating computer-generated domain-oriented vocabularies. Inf. Process. Manag. **26**(6), 791–801 (1990)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
8. Erjavec, T., Fišer, D., Ljubešić, N.: The KAS corpus of slovenian academic writing. Lang. Resour. Eval. **55**(2), 551–583 (2021)
9. Fišer, D., Suchomel, V., Jakubček, 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)
10. Frantzi, K.T., Ananiadou, S., Tsujii, J.: The *C-value/NC-value* method of automatic recognition for multi-word terms. In: Nikolaou, C., Stephanidis, C. (eds.) ECDL 1998. LNCS, vol. 1513, pp. 585–604. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-49653-X_35
11. Gao, Y., Yuan, Yu.: Feature-less end-to-end nested term extraction. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11839, pp. 607–616. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6_55

12. 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)
13. 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
14. Jemec Tomazin, M., Trojar, M., Žagar, M., Atelšek, S., Fajfar, T., Erjavec, T.: Corpus of term-annotated texts RSDO5 1.0 (2021)
15. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. *Nat. Lang. Eng.* **1**(1), 9–27 (1995)
16. 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)
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. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. arXiv preprint [arXiv:1909.11942](https://arxiv.org/abs/1909.11942) (2019)
20. 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)
21. Le, H., et al.: FlauBERT: unsupervised language model pre-training for French. In: LREC (2020)
22. 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. *Terminol. Int. J. Theor. Appl. Issues Special. Commun.* **16**(1), 77–106 (2010)
23. Lingpeng, Y., Donghong, J., Guodong, Z., Yu, N.: Improving retrieval effectiveness by using key terms in top retrieved documents. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 169–184. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31865-1_13
24. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692) (2019)
25. Ljubešić, N., Fišer, D., Erjavec, T.: KAS-term: extracting slovene terms from doctoral theses via supervised machine learning. In: Ekštejn, K. (ed.) TSD 2019. LNCS (LNAI), vol. 11697, pp. 115–126. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27947-9_10
26. Maldonado, A., Lewis, D.: Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In: Proceedings of The 12th International Conference on Terminology and Knowledge Engineering, pp. 91–100 (2016)
27. Martin, L., et al.: Camembert: a tasty French language model. arXiv preprint [arXiv:1911.03894](https://arxiv.org/abs/1911.03894) (2019)
28. Pavlopoulos, J., Androutsopoulos, I.: 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), pp. 44–52 (2014)

29. Jagarlamudi, J., Daumé, H.: Extracting multilingual topics from unaligned comparable corpora. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 444–456. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12275-0_39
30. Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., Pollak, S.: TermEnsembler: an ensemble learning approach to bilingual term extraction and alignment. *Terminol. Int. J. Theoretical Appl. Issues Special. Commun.* **25**(1), 93–120 (2019)
31. 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)
32. Rigouts Terryn, A., Hoste, V., Lefever, E.: In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Lang. Resour. Eval.* **54**(2), 385–418 (2020)
33. Rigouts Terryn, A., Hoste, V., Lefever, E.: HAMLET: Hybrid adaptable machine learning approach to extract terminology. *Terminol.* (2021)
34. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT: a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint [arXiv:1910.01108](https://arxiv.org/abs/1910.01108) (2019)
35. Tran, H.T.H., Martinc, M., Doucet, A., Pollak, S.: Can cross-domain term extraction benefit from cross-lingual transfer? In: International Conference on Discovery Science, pp. 363–378. Springer (2022)
36. 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: Submitted to Slovenian conference on Language Technologies and Digital Humanities (2022, under review)
37. Hanh, T.T.H., Doucet, A., Sidere, N., Moreno, J.G., Pollak, S.: Named entity recognition architecture combining contextual and global features. In: Ke, H.-R., Lee, C.S., Sugiyama, K. (eds.) ICADL 2021. LNCS, vol. 13133, pp. 264–276. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91669-5_21
38. Vaswani, A., et al.: Attention is all you need. arXiv preprint [arXiv:1706.03762](https://arxiv.org/abs/1706.03762) (2017)
39. Vintar, S.: Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminol. Int. J. Theoretical Appl. Issues Specialized Commun.* **16**(2), 141–158 (2010)
40. 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)
41. Yang, J., et al.: Towards making the most of BERT in neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9378–9385 (2020)
42. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. arXiv preprint [arXiv:1906.08237](https://arxiv.org/abs/1906.08237) (2019)
43. Zhang, Z., Gao, J., Ciravegna, F.: Semre-rank: improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. *ACM Trans. Knowl. Discov. Data (TKDD)* **12**(5), 1–41 (2018)