Machine Learning
https://doi.org/10.1007/s10994-023-06506-7

Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling?

Hanh Thi Hong Tran[1,2,3] · Matej Martinc[2] · Andraz Repar[2] · Nikola Ljubešić[2] · Antoine Doucet[3] · Senja Pollak[2]

Received: 5 March 2023 / Revised: 12 August 2023 / Accepted: 16 December 2023
© The Author(s) 2024

Abstract
Automatic term extraction (ATE) is a natural language processing task that eases the effort of manually identifying terms from domain-specific corpora by providing a list of candidate terms. In this paper, we treat ATE as a sequence-labeling task and explore the efficacy of XLMR in evaluating cross-lingual and multilingual learning against monolingual learning in the cross-domain ATE context. Additionally, we introduce NOBI, a novel annotation mechanism enabling the labeling of single-word nested terms. Our experiments are conducted on the ACTER corpus, encompassing four domains and three languages (English, French, and Dutch), as well as the RSDO5 Slovenian corpus, encompassing four additional domains. Results indicate that cross-lingual and multilingual models outperform monolingual settings, showcasing improved F1-scores for all languages within the ACTER dataset. When incorporating an additional Slovenian corpus into the training set, the multilingual model exhibits superior performance compared to state-of-the-art approaches in specific scenarios. Moreover, the newly introduced NOBI labeling mechanism enhances the classifier's capacity to extract short nested terms significantly, leading to substantial improvements in Recall for the ACTER dataset and consequentially boosting the overall F1-score performance.

Keywords Term extraction · XLMR · Sequence labeling · Cross-lingual · Cross-domain · Nested terms

Editors: Dino Ienco, Roberto Interdonato, Pascal Poncelet.

✉ Senja Pollak
senja.pollak@ijs.si

Hanh Thi Hong Tran
tran.hanh@ijs.si

[1] Jožef Stefan International Postgraduate School, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[2] Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia

[3] University of La Rochelle, 23 Av. Albert Einstein, La Rochelle, France

Published online: 27 March 2024

1 Introduction

Terms are textual expressions that denote concepts in a specific field of expertise. They are beneficial for several terminographical tasks performed by linguists (e.g., construction of specialized terminological dictionaries (Le Serrec et al., 2010)). Moreover, terms can also support and improve several downstream natural language processing (NLP) tasks (e.g., topic detection (ElKishky et al., 2014), information retrieval (Lingpeng et al., 2005), machine translation (Wolf et al., 2011)). To ease the time and effort needed to manually identify terms in domain-specific corpora, automatic term extraction (ATE) approaches were proposed.

The TermEval 2020 shared task, organized as part of the CompuTerm workshop (Rigouts et al., 2020a), presented one of the first opportunities to systematically study and compare several ATE architectures with the introduction of the Annotated Corpora for Term Extraction Research (ACTER) dataset (Rigouts et al., 2020a, b). While the workshop was a significant step forward in systematic comparison, the less-resourced languages (e.g., Slovenian) have not yet been sufficiently explored and remain a research gap. Furthermore, there is still room for improvement in performance. In our previous study (Tran et al., 2022a), the conducted error analysis pointed out that the two most common errors that the tested classifiers made were to predict a shorter term nested in the ground truth term and vice versa, i.e., the model sometimes generates the terms not covered in the ground truth, containing a nested term. This insight leads to a hypothesis about the insufficiency of the widely used BIO labeling regime (Hazem et al., 2020). This regime does not allow labeling the nested terms and giving the model the necessary information to avoid the above mistakes.

Inspired by the success of Transformers (Hazem et al., 2020) and the rise of cross-lingual learning (Lang et al., 2021), our research delves into the effectiveness of the XLMR (Conneau et al., 2020) in multilingual and cross-lingual scenarios. First, having a single model that works across several languages is important, as it can be used also in the languages not seen during the training. Instead of having to construct language-specific models, multilingual and cross-lingual models can be directly used on any new language that is supported by XLMR. In addition, for the languages where the data is available, having a single model instead of many language-specific models is a much simpler solution, and can also make the models less dataset-specific.

Our approach frames the ATE task as a sequence-labeling problem, as this strategy has proven successful in various NLP tasks like Named Entity Recognition (NER) (Lample et al., 2016; Tran et al., 2021) and Keyword Extraction (Martinc et al., 2021). Additionally, we extend our previous work (Tran et al., 2022a) by introducing an innovative nested term labeling mechanism, incorporating two extra labels for single nested terms, and rigorously evaluating the model's performance in cross-lingual and multi-lingual settings. This comprehensive exploration showcases the power of a multilingual pretrained language model with cross-lingual and multi-lingual settings in capturing and understanding diverse linguistic nuances. The experiments are conducted in the cross-domain setting on the ACTER dataset[1] containing texts in four domains (Corruption, Wind energy, Equitation, and Heart failure) with three languages (English, French, and Dutch) and the RSDO5 corpus[2] (Jemec

[1] https://github.com/AylaRT/ACTER.

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

Tomazin et al., 2021) containing Slovenian texts from four domains (Biomechanics, Chemistry, Veterinary, and Linguistics).

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

- We propose a new NOBI annotation mechanism to better capture single nested terms. When a dataset contains a relevant proportion of nested terms, the new labeling regime improves the Recall of the models by a large margin, leading also to further improvements in the F1-score. This is also the main novelty compared to the shorter conference version (Tran et al., 2022a) of this paper.
- We systematically evaluate the performance of the XLMR on the cross-domain term extraction task in two datasets covering English, French, Dutch, and a less-resourced Slovenian in both standard BIO and the novel NOBI scheme.
- We compare the performance among cross-lingual, multilingual, and monolingual approaches to determine the general applicability of multilingual language models for sequence labeling in both rich- and less-resourced languages. The datasets using BIO and NOBI annotation regimes are both considered.

2 Related work

The history of ATE has its beginnings during the 1990s with research done by Damerau (1990) and Justeson and Katz (1995). ATE systems usually employ the two-step procedure: (1) extracting a list of candidate terms, and (2) determining which candidate terms are correct.

2.1 Approaches based on term characteristics

Traditional approaches relied on distinctive linguistic aspects of terms to extract possible candidates. Several NLP tools (e.g., tokenization, lemmatization, stemming, PoS tagging) are employed to obtain linguistic profiles of term candidates. As a heavily language-dependent approach, the better the quality of the pre-processing tools (e.g., FLAIR (Akbik et al., 2019), Stanza (Qi et al., 2020)), the better the quality of linguistic methods. More recent studies preferred the statistical approach, which commonly relies on the assumption that a higher candidate term frequency in a domain-specific corpus implies a higher likelihood that a candidate is an actual term. Some measures relying on this assumption include termhood (Vintar, 2010), unithood (Daille et al., 1994) or C-value (Frantzi et al., 1998). More popular statistical approaches also considered the frequency of the term internal words compared to the term frequency to identify rare terms and remove frequent words. Many current systems still apply this approach's variation, or hybrid mechanisms that combine linguistic and statistical information (Kessler et al., 2019; Repar et al., 2019).

2.2 Approaches based on machine learning and deep learning

Recent advances in representation learning and deep neural networks have also influenced term extraction. Several embedding techniques have been investigated for the task at hand, e.g., uni-gram (Amjadian et al., 2016), non-contextual (Zhang et al., 2018), contextual (Kucza et al., 2018) word embeddings, and the hybrid ones (Gao & Yuan, 2019). The first use of language models for the ATE task was in the TermEval 2020 (Rigouts et al., 2020a)

where the winning approach on the Dutch corpus used BiLSTM-based neural architecture with GloVe embeddings while the winning solution on the English corpus (Hazem et al., 2020) extracted all possible n-gram combinations, which are then fed into a BERT binary classifier that determines for each n-gram inside a sentence, whether it is a term. Besides, several Transformer-based variations have also been investigated (e.g., RoBERTa, CamemBERT (Hazem et al., 2020)). Further work includes HAMLET by Terryn et al., 2021, which proposes a hybrid adaptable machine learning classifier that combines linguistic and statistical clues to detect terms.

Recently, sequence-labeling and cross-lingual approaches toward ATE have been gaining traction. Kucza et al. (2018) was one of the first to model term extraction as a sequence-labeling task. Cross-lingual sequence labeling was, on the other hand, explored in Conneau et al. (2020), Lang et al. (2021), Hazem et al. (2022), and Tran et al. (2022a), who took advantage of XLMR, the model we also employ in this work. Lang et al. (2021) compared different cross-lingual approaches, including a sequence classifier, and a token classifier on this sequence-labeling task, and further proposed a sequence-to-sequence (seq2seq) approach, which used mBART (Liu et al., 2020) to generate sequences of comma-separated terms from the input. The results demonstrate the capability of multilingual models to outperform monolingual ones in some specific scenarios and the potential of cross-lingual learning.

Finally, in our conference paper (Tran et al., 2022a) that we extend in this journal paper, we leveraged the multilingual setup by fine-tuning the model using training datasets from several languages and then applying the model to their test sets, separately. By doing so, we examined whether adding more data from other languages to the training set that matches the target language in the testing set improves the model's predictive performance. After adding the Slovenian corpus into the ACTER training set, our multilingual model demonstrated a significant improvement in Recall across all test languages compared to the monolingual one.

2.3 Approaches for Slovenian term extraction

For Slovenian, the language used in our study, and for less-resourced languages in general, the research is still hindered by the lack of gold standard corpora and limited use of neural methods. Things are nevertheless slowly improving. In recent years, the Slovenian KAS corpus was compiled (Erjavec et al., 2021), quickly followed by another corpus designed for term extraction, the RSDO5 corpus.[3] Regarding the methods, Vintar (2010) was one of the first to propose statistical approaches for Slovenian ATE tasks. After that, Ljubešić et al. (2019) introduced a hybrid one, in which they extract the initial candidate terms using the CollTerm tool (Pinnis et al., 2019), a rule-based system employing a complex language-specific set of term patterns from the Slovenian SketchEngine (Fišer et al., 2016). Meanwhile, Repar et al. (2019) focuses on term extraction and alignment, where the novelty is the evolutionary algorithm for the term alignment.

The deep neural approaches have not been sufficiently explored for Slovenian data yet. The only neural approach towards Slovenian ATE was proposed in our recent study (Tran et al., 2022b). There, we implemented the Transformers-based sequence-labeling approach, which we extend in this study, in a cross-lingual and multilingual evaluation. Another

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

problem is that often no open-sourced code is available for most current benchmark systems, hindering their reproducibility (for Slovenian, only the code from Ljubešić et al. (2019) and Tran et al. (2022b) methods are available).

2.4 Extraction of nested terms

In many practical applications, it is common that the terms have a nested structure where a term could contain other terms or be part of others. Vintar (2004) first suggested ranking and/or discarding nested terms using the C-value, but their results were unsatisfactory. Marciniak and Mykowiecka (2015) later identified them by combining grammatical correctness and normalized pointwise mutual information (NPMI) based on bigrams in a corpus. However, this method's efficiency relies heavily on corpus features (e.g., size, thematic homogeneity, and phrase frequency). Recently, Gao and Yuan (2019) proposed an end-to-end architecture that employs classification and ranking for n-gram candidates in text sequences. Nonetheless, this suffers from reduced Recall due to ranking and its threshold output is not applicable to new, unseen domains. Since then, no further methodologies have been proposed, leaving a gap in extracting nested terms for term extraction tasks.

Regarding other NLP downstream tasks sharing the same mechanisms (e.g., NER, Keyword Extraction), besides the common sequence tag schemes (e.g., BIO (Ramshaw & Marcus, 1999), IOBES (Lester, 2020), BMEWO (Ratinov & Roth, 2009), BILOU (Ratinov & Roth, 2009)) for both flat and nested ones, we can categorize the methods to capture nested entities into four main types: (1) sequence labeling, (2) hypergraph-based, (3) sequence-to-sequence (Seq2Seq), and (4) span-based methods. However, none of them except the BIO regime for the sequence-labeling approach has been applied for term extraction yet.

3 Methodology

Section 3.1 presents a brief description of our chosen datasets. We demonstrate the general methodology, experimental setup, and implementation details in Sects. 3.2 and 3.3. Finally, in Sect. 3.4, we present our chosen evaluation metrics.

3.1 Datasets

The experiments were conducted on ACTER (Rigouts et al., 2020a) and RSDO5 version 1.1 (Jemec Tomazin et al., 2021), both comprising texts from diverse languages and domains. The ACTER dataset is a collection of 12 corpora covering four domains (Corruption (corp), Equitation (equi), Wind energy (wind), and Heart failure (htfl)) in three languages (English (en), French (fr) and Dutch (nl)). The dataset has two types of gold standard annotations: one containing both terms and named entities (NES), and the other one containing only terms (ANN). The second dataset is the RSDO5 version 1.1 (Jemec Tomazin et al., 2021), which contains texts in Slovenian (sl), a less-resourced Slavic language with rich morphology. The corpus contains 12 documents collected from 2000 to 2019 covering domains of Biomechanics (bim), Chemistry (kem), Veterinary (vet), and Linguistics (ling). The data analysis is depicted in Figs. 14, 15 , 16, 17 and 18 and Table 7.

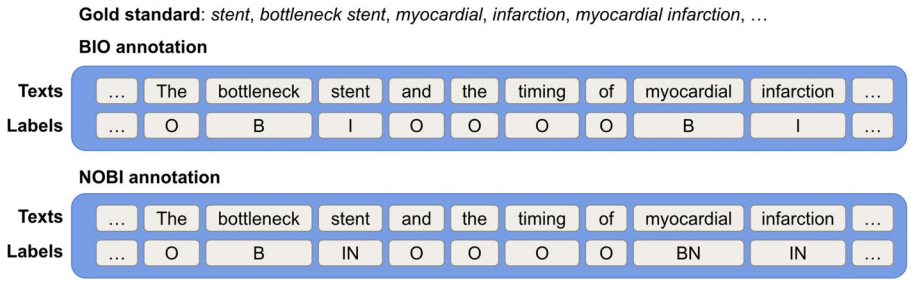

Fig. 1 An example of BIO and NOBI annotation regimes in the ACTER corpus

3.2 Experimental setup

We consider ATE as a sequence-labeling task where the model returns a label for each token in a text sequence using two different labeling regimes: the benchmark BIO labeling scheme (Lang et al., 2021; Rigouts et al., 2021) and our novel annotation scheme called NOBI. In the BIO regime, B stands for the beginning word in the term, I stands for the word inside the term, and O stands for the word not part of the term. The terms from a gold standard list are first mapped to the tokens in the raw text and each word inside the text sequence is annotated with one of three labels (see the upper example in Fig. 1). However, it is not optimized for nested term extraction. Thus, we propose NOBI, an annotation regime with two additional labels BN and IN, referring to a word being in the beginning or inside the nested term, respectively (see the lower example in Fig. 1). An annotation regime with two additional labels BN and IN, where N refers to nested single-word terms, which can be at the beginning (BN) or inside (IN) position of a longer term.

In Fig. 1, the gold standard contains the following terms: *"stent", "bottleneck stent", "myocardial", "infarction", "myocardial infarction"*, etc. In the BIO regime, we ignore the single nested terms, thus, we only mark *"bottleneck"* as the beginning (B) and *"stent"* as the inside (I) of the full term *"bottleneck stent"*. Similarly, *"myocardial"* is the beginning (B), and *"infarction"* is the inside (I) of the full term *"myocardial infarction"*. However, in the NOBI regime, we consider *"bottleneck stent"* and *"stent"* as two different terms where *"stent"* is the nested term of *"bottleneck stent"*, in contrast to the BIO scheme, where the model extracts just the *"bottleneck stent"* as a term. Similarly, *"myocardial"* and *"infarction"* are two separate terms that are nested in *"myocardial infarction"*. Therefore, an additional label N is added to the label of *"stent"*, *"myocardial"*, and *"infarction"*.

We do not consider either multi-word nested terms or terms nested in other nested terms – so-called nested terms on the second or higher levels – due to their rarity in the corpora and gold standards (see the nested frequency in the gold standard from Figs. 16 and 18 in Appendix). Despite the difference in the number of terms in each language and domain, the percentage of unique nested terms in all languages and domains is somewhat consistent, ranging around one-third of the total unique terms in the gold standards. However, the number of terms nested in other nested terms only takes one-tenth and one-twelfth of the total amount of unique terms in both corpora, respectively, and the amounts are even much smaller if we specify the ratio per nested level (e.g., in the second level, third level). We also demonstrate in Table 7 in Appendix the proportion of the nested terms with different

word lengths k where $k = \{1, 2, 3, 4, \geq 5\}$ for each domain and language of both corpora. The last column on the right calculates the percentage of single-word nested terms in total nested terms in the first level. On average, the amount of single-word nested terms accounts for 78.06% above all the nested terms on the first levels in the corpora. Therefore, we only label single-word nested terms on the first level.

For both labeling regimes, we experiment with XLMR, a Transformer-based model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. The model is first trained to predict a label for each token in the input text sequence (e.g., we model the task as token classification) and then applied to the unseen text (test data). Finally, from the tokens or token sequences labeled as terms, the final candidate term list for the test data is composed. Note that when the NOBI annotation regime is used, the terms labeled with BN and IN are added to the final term list separately, together with the terms in which they are nested.

We evaluate the cross-domain performance of XLMR in monolingual, cross-lingual, and multilingual settings. Altogether, 78 different scenarios per annotation regime are tested. The distinct settings are described below.

1. **Monolingual setup.** We evaluate how well the model performs when there is a language-specific training corpus available and there is a match between the language of the train set and the language of the test set. For better comparison with other existing approaches, we apply the same configuration as in the TermEval 2020 shared task where Heart failure of each language is considered as the test set. Thus, we fine-tune our model on a single language, which means we train three monolingual models for three languages (English, French, Dutch) and test each model in the same language for each annotation regime. Besides, we train 12 monolingual models for each annotation regime for Slovenian given 12 different combinations of train-validation-test split regarding the domains.
2. **Cross-lingual setup.** We evaluate the capability of the model to apply the knowledge learned in one or more languages for ATE in another unseen language. Therefore, we fine-tune the ATE model on one or more languages (e.g., English and Dutch) and test it on another language not appearing in the train set (e.g., French). In this scenario, we, therefore, examine how well the model performs without the language-specific training corpus and how good the knowledge transfer between different languages is.
3. **Multilingual setup.** We fine-tune our model using a.) training datasets from the languages in the ACTER dataset (English, French, and Dutch) or b.) training datasets from the languages in the ACTER dataset plus the Slovenian training dataset from the RSDO5 corpus, and then apply the model to the test sets of all languages in the ACTER dataset. By doing so, we examine whether adding more data from other languages to the training set in the target language improves the predictive performance of the model.

All three settings are applied in a cross-domain evaluation scenario, where we use two domains for training, another domain for validation, and the rest for testing. One exception is the multilingual and cross-lingual settings with the additional Slovenian corpus in the training set, where we use two domains from ACTER corpora and all domains from the RSDO5 corpus for the training. This way, we can evaluate the model's generalization capabilities to adapt knowledge in one or more domains to a new, unseen arbitrary one and, therefore, much more useful. In the ACTER dataset, we use the Corruption and Wind energy domains for training, the Equitation domain for validation, and the Heart

failure domain for testing, in order to allow for a direct comparison with other benchmark approaches from the related work, which employ the same train-validation-test setting (Lang et al., 2021). Meanwhile, in the RSDO5 corpus, we explore different train-validation-test combinations.

We divide the dataset into train-validation-test splits. The model is fine-tuned on the training set to predict the probability for each word in a word sequence whether it is a part of the term (B, I), whether it is a nested term (BN for nested terms at the beginning of a multi-word term, IN for nested terms at non-beginning positions of a multi-word term), or not part of the term (O). To do that, an additional token classification head containing a feed-forward layer with a softmax activation is added on top of each model.

3.3 Implementation details

We employ the XLMR token classifier from Huggingface.[4] We fine-tune the model for up to 20 epochs (i.e., the early stopping regime via the validation set) using the learning rate of 2e-05, training and evaluation batch size of 32, and sequence length of 512 tokens, since this hyperparameter configuration performed the best on the validation set. First, the documents are split into sentences. Then, the sentences with more than 512 tokens are truncated, while those with less than 512 tokens are padded with a special $< PAD >$ token at the end.

During fine-tuning, the model is evaluated on the validation set after each training epoch, and the best-performing model is applied to the test set. Note that the model with BIO annotation regime will predict the probability of whether the word is a part of the terms (B, I) or not (O) while the one with NOBI regime will predict the probability in the same manner as BIO except for the additional information on the nested terms of the first level (BN, IN) where each word with the N label will be considered as an individual single-word candidate term. The sequences identified as terms are extracted from the text and put into a set of all predicted candidate terms. A post-processing step to lowercase all candidate terms is applied before we compare our derived candidate list with the gold standard.

3.4 Evaluation metrics

We evaluate the performance of the ATE systems by comparing the candidate list extracted from the test set with the manually annotated gold standard term list for that specific test set. We use exact string matching to compare the retrieved terms to the ones in the gold standard and calculate Precision (P), Recall (R), and F1-score (F1). These evaluation metrics have also been used in related work (Hazem et al., 2020; Lang et al., 2021; Rigouts et al., 2020a; Ljubešić et al., 2019), therefore, our results are directly comparable to the benchmarks.

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

Table 1 Evaluation on the ACTER dataset given Heart failure as a test set. For each test set, bold is used to indicate the best model in terms of P, R, and F1 for each test set (ANN and NES) and each annotation scheme separately (BIO and NOBI). The arrows are used for the comparison of BIO and NOBI for each setting, where ↑ is used to show the better performance of NOBI compared to BIO, while ↓ denotes the lower performance of NOBI compared to BIO. In blue, we indicate the best model in terms of the F1 for each test set

Train language	ANN						NES					
	BIO			NOBI			BIO			NOBI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
English test set.												
en	58.1	48.1	52.6	↓ 57.5	↑ 48.6	↑ 52.7	62.1	52.1	56.7	↓ 58.6	↑ 55.2	↑ 56.9
fr	56.9	33.2	42.0	↓ 54.2	↑ 34.7	↑ 42.3	60.0	39.1	47.4	↓ 57.8	↑ 44.3	↑ 50.2
nl	55.6	56.4	**56.0**	↑ 57.6	↑ 58.4	↑ **58.0**	54.4	57.7	56.0	↑ 56.9	↑ 61.2	↑ 59.0
fr, sl	47.1	65.8	54.9	↓ 42.5	↑ 68.8	↓ 52.5	49.2	64.3	55.7	↓ 44.6	↑ 66.6	↓ 53.4
nl, sl	45.7	**66.3**	54.1	↑ 46.0	↑ 67.8	↑ 54.8	48.1	65.4	55.5	↑ 49.2	↑ 67.0	↑ 56.8
fr, nl	**60.8**	46.8	52.9	↓ 57.5	↓ 41.5	↓ 48.2	62.3	50.5	55.7	↓ 58.6	↑ 52.0	↑ 55.1
fr, nl, sl	50.0	62.4	55.5	↓ 48.3	↑ 67.2	↑ 56.2	52.1	63.2	57.2	↓ 49.5	↑ 65.3	↓. 56.3
en, fr	57.2	51.2	54.0	↑ **58.0**	51.2	↑ 54.4	60.4	51.5	55.6	↓ **59.5**	↑ 54.2	↑ 56.7
en, nl	58.0	48.7	52.9	↓ 54.0	↑ 56.1	↑ 55.0	**62.4**	51.4	56.4	↓ 57.4	↑ 58.6	↑ 58.0
en, sl	48.1	63.2	54.6	↑ 49.0	↑ 65.7	↑ 56.1	54.9	63.8	59.0	↓ 50.8	↑ 64.4	↓ 56.8
en, fr, sl	48.1	64.2	55.0	↑ 51.1	↑ 67.2	↑ **58.0**	58.4	61.1	**59.7**	↓ 55.2	↑ 63.4	↓ 59.0
en, nl, sl	48.4	65.0	55.4	↓ 44.8	↑ 68.6	↓ 54.2	54.5	63.3	58.6	↓ 53.1	↑ 67.3	↑ **59.3**
en, fr, nl	56.8	53.0	54.9	↓ 55.7	↓ 51.0	↓ 53.3	60.8	52.6	56.4	↓ 57.4	↑ 59.8	↑ 58.6
en, fr, nl, sl	45.9	**66.3**	54.2	↓ 45.5	↑ **69.3**	↑ 54.9	48.3	**65.7**	55.6	↑ 51.9	↑ **68.4**	↑ 59.0
cross-ling. avg.	52.7	55.2	52.6	↓ 51.0	↑ 56.4	↓ 52.0	54.4	56.7	54.6	↓ 52.8	↑ 59.4	↑ 55.1
multi-ling. avg.	51.8	58.8	54.4	↓ 51.2	↑ 61.3	↑ 55.1	57.1	58.5	57.3	↓ 55.0	↑ 62.3	↑ 58.2
French test set.												
fr	**70.5**	44.4	54.5	↓ 66.3	↑ 48.9	↑ 56.3	**72.4**	48.5	58.1	↓ 65.9	↓ 54.7	↑ 59.8
en	66.7	47.9	55.8	↑ 67.8	↓ 44.8	↓ 53.9	70.6	53.8	61.1	↓ 65.3	↓ 53.3	↓ 58.7
nl	66.5	51.5	58.0	↓ 64.7	↓ 47.0	↓ 55.1	67.6	53.2	59.5	↓ 67.5	↓ 53.1	↓ 59.4
en, sl	60.2	61.4	60.8	↑ 61.1	↓ 57.5	↓ 59.2	57.8	**62.5**	60.1	↑ 62.9	↓ 56.0	↓ 59.2
nl, sl	61.4	60.4	60.9	↓ 59.5	↓ 58.5	↓ 59.0	61.8	59.9	60.8	↑ 63.1	↓ 56.7	↓ 59.7
en, nl	65.3	44.2	52.7	↓ 65.2	↑ 47.9	↑ 55.2	68.7	52.4	59.4	↑ **69.3**	↓ 50.6	↓ 58.5
en, nl, sl	58.7	61.0	59.8	↓ 55.3	↑ 63.2	↓ 59.0	60.9	62.0	61.5	↓ 59.0	↑ 62.3	↓ 60.6
fr, en	63.7	52.4	57.5	↑ 65.7	↓ 49.5	↓ 56.4	68.1	52.8	59.5	↓ 67.2	↓ 49.6	↓ 57.1
fr, nl	69.2	48.3	56.9	↓ **66.4**	↑ 48.4	↓ 56.0	70.7	49.5	58.3	↓ 66.1	↑ 54.2	↑ 59.6
fr, sl	65.0	56.6	60.5	↓ 58.8	↑ 62.3	60.5	65.3	57.6	61.2	↓ 56.9	↑ 64.0	↓ 60.2
fr, en, sl	61.5	58.6	60.0	↑ 63.2	↑ 60.5	↑ **61.8**	67.4	57.5	**62.1**	↓ 64.1	↑ 61.6	↑ **62.9**
fr, nl, sl	64.9	58.2	**61.4**	↓ 61.5	↑ 61.0	↓ 61.3	65.3	57.9	61.4	↓ 63.1	↑ **62.7**	↑ **62.9**
fr, en, nl	68.0	50.7	58.1	↓ 65.4	↓ 46.9	↓ 54.6	70.2	52.1	59.8	↓ 63.8	↑ 56.5	↑ 60.0
en, fr, nl, sl	58.1	**61.6**	59.8	↑ 60.3	↑ **62.8**	↑ 61.6	59.5	**62.5**	61.0	↑ 64.2	↓ 59.5	↑ 61.7
cross-ling. avg.	63.1	54.4	58.0	↓ 62.3	↓ 53.3	↓ 56.9	64.6	57.3	60.4	↓ 64.5	↓ 55.3	↓ 59.4
multi-ling. avg.	64.3	55.2	59.2	↓ 63.0	↑ 55.9	↓ 58.9	66.6	55.7	60.5	↓ 63.6	↑ 58.3	↑ 60.6
Dutch test set.												
nl	70.3	62.2	66.0	↑ 71.2	↑ 64.1	↑ 67.5	73.3	61.5	66.9	↑ **73.5**	↑ 62.6	↑ 67.6
en	69.2	61.1	64.9	↑ 71.0	61.1	↑ 65.7	73.0	63.0	67.6	↓ 69.4	↑ 68.4	↑ 68.9
fr	72.1	51.0	59.8	↓ 70.4	↑ 55.6	↑ 62.2	73.6	55.5	63.3	↓ 70.4	↑ 62.4	↑ 66.2
en, sl	59.5	76.6	67.0	↑ 61.6	↑ 78.3	↑ 68.9	61.1	73.6	66.7	↑ 61.6	↑ **75.7**	↑ 68.4
fr, sl	62.5	74.7	68.1	↓ 58.7	↑ **79.3**	↓ 67.5	61.6	71.2	66.1	↓ 59.4	↑ 75.1	↑ 66.3
en, fr	72.5	61.7	66.7	↓ 70.8	↓ 60.1	↓ 65.0	73.1	63.5	68.0	↓ 72.5	↓ 61.2	↓ 66.4
en, fr, sl	59.6	**77.0**	67.2	↑ 61.1	↑ 78.2	↑ 68.6	66.6	69.6	68.1	↓ 66.4	↑ 74.9	↑ 70.4
nl, en	69.3	60.2	64.4	↓ 68.6	↑ 62.7	↑ 65.5	74.4	61.7	67.4	↓ 70.7	↑ 66.3	↑ 68.4
nl, fr	**75.7**	56.7	64.8	↓ **73.2**	↑ 58.1	64.8	**76.7**	59.6	67.1	↓ 73.0	↑ 60.6	↓ 66.2
nl, sl	65.8	72.7	**69.1**	↓ 65.0	↑ 77.0	↑ 70.5	69.9	69.7	**69.8**	↓ 68.6	↑ 72.5	↑ 70.5
nl, en, sl	64.7	73.0	68.6	↓ 60.0	↑ 80.6	↑ 68.8	68.7	70.3	69.5	↓ 67.6	↑ 74.2	↑ 70.8
nl, fr, sl	69.2	69.0	**69.1**	↓ 65.2	↑ 76.5	↑ 70.4	69.4	69.4	69.4	↓ 65.4	↑ 74.4	↑ 69.6
nl, en, fr	69.9	64.3	67.0	↑ 72.1	↓ 55.5	↓ 62.7	73.7	62.9	67.9	↓ 71.1	↑ 64.9	↓ 67.8
en, fr, nl, sl	62.7	75.5	68.5	↑ 64.5	↑ 78.1	↑ **70.6**	63.6	**73.7**	68.3	↑ 69.2	↓ 73.2	↑ **71.1**
cross-ling. avg.	65.9	67.0	65.6	↓ 65.6	↑ 68.8	↑ 66.3	68.2	66.1	66.6	↓ 66.6	↑ 69.6	↑ 67.8
multi-ling. avg.	68.2	67.3	67.4	↓ 66.9	↑ 69.8	↑ 67.6	70.9	66.8	68.5	↓ 69.4	↑ 69.4	↑ 69.2

4 Results

In this Section, we determine the predictive power of monolingual, cross-lingual, and multilingual learning in ACTER and RSDO5 test sets as well as compare the results from our proposed approaches to the SOTA from the related work.

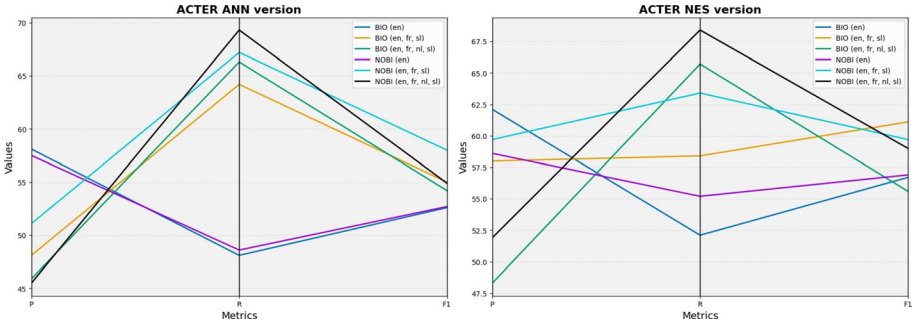

Fig. 2 Parallel Coordinates Plot in performance of XLMR classifier for the English test set

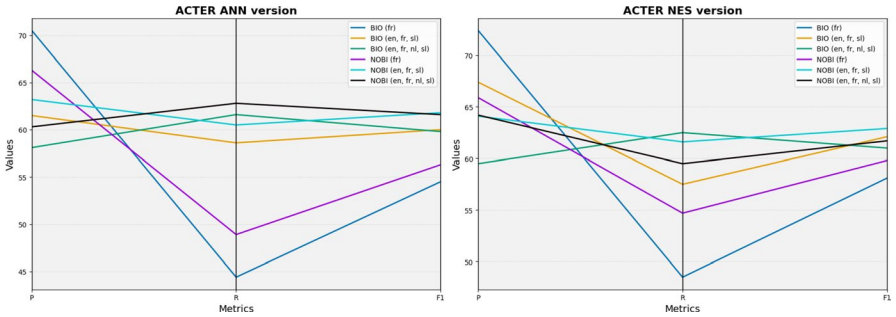

Fig. 3 Parallel Coordinates Plot in performance of XLMR classifier for the French test set

4.1 Results on the ACTER test set

The performance of the XLMR classifier regarding P, R, and F1 on the ACTER test set using BIO and NOBI annotation regimes are presented in Table 1. The comparison between BIO and NOBI is indicated with arrows, where ↑ is used to show better performance of NOBI in the same setting, while ↓ denotes lower performance. No matter which annotation scheme, the results indicate that the cross-lingual and multilingual models in both versions of test data, where one excludes the named entities of the test data (ANN) and the other includes them (NES), tend to surpass the performance of the monolingual ones according to all evaluation metrics, except for the Precision obtained by the French monolingual model on the French test set when the BIO scheme is used and Dutch monolingual model on the Dutch test set when NOBI scheme is used.

Multilingual models tend to outperform cross-lingual ones in F1. However, multilingual models have a tendency to lose their competency in Precision toward monolingual and cross-lingual ones. By adding the Slovenian corpus with four different domains into the training set, the multilingual model demonstrates a significant improvement in Recall across all test languages compared with the monolingual setting. It also outperforms other models in the F1 when we evaluate it in all three test sets in both annotation schemes. However, this improvement is at the cost of Precision.

When it comes to the comparison of the two annotation regimes, using the NOBI annotations in many cases improves the Recall of the model. This is especially visible in

Table 2 F1 comparison between our XLMR classifier in multilingual settings and related work in ACTER corpora

Methods	English		French		Dutch	
	ANN	NES	ANN	NES	ANN	NES
Winning teams (Hazem et al. 2020)	45.0	46.7	45.9	48.2	18.6	18.7
HAMLET (Rigouts et al. 2021)	54.2	55.4	60.2	60.8	66.1	66.0
Sequence Classifier (Lang et al. 2021)	x	46.0	x	48.1	x	58.0
NMT (Lang et al. 2021)	x	55.3	x	57.6	x	59.6
Token classifier (Lang et al. 2021)	x	58.3	x	57.6	x	69.8
NMF-based approaches (Nugumanova et al. 2022)	33.5	33.7	30.9	30.7	30.1	30.3
BIO classifier	54.9	**59.7**	61.4	62.1	69.1	69.8
NOBI classifier	**58.0**	59.3	**61.8**	**62.9**	**70.6**	**71.1**

Bold indicates the best result for each test set

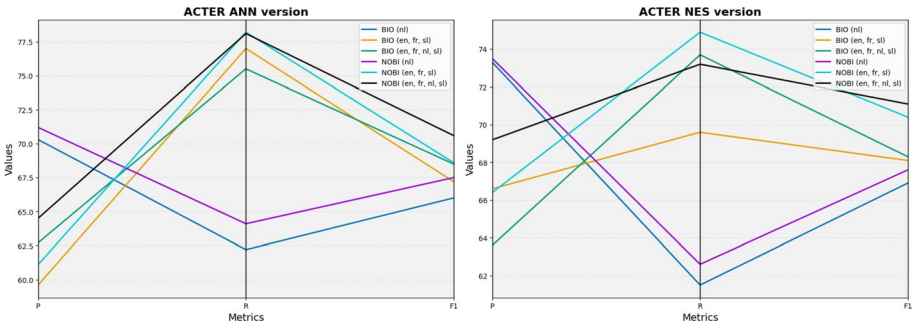

Fig. 4 Parallel Coordinates Plot in performance of XLMR classifier for the Dutch test set

the monolingual and multilingual settings (see Figs. 2, 3, and 4) in which the models are trained in multiple languages including the language of the test sets for all scenarios, and cross-lingual settings in which the models are trained on just one language and applied to the others except for French test set. A substantial increase in Recall also tends to lead to the improvement of the overall F1.

The best models from our combinations include: (1) For the English and French test sets, the best results were obtained with English, French, and Slovenian training data; and (2) For the Dutch test set, the best results were gained with the multilingual classifiers of all four languages. Thus, we compare the multilingual XLMR classifier fine-tuned on the pre-defined test language and multiple languages (trained in at least three languages including Slovenian and the test set's language) using the ACTER dataset in both annotation regimes. This showcases the power of a multilingual pretrained language model with multilingual settings - using (1) English, French, and Slovenian; and (2) all four languages as the training set - in capturing and understanding diverse linguistic nuances in comparison with a monolingual one. Additionally, the NOBI regime outperforms BIO ones for most of the testing scenarios.

Besides, we also compare the proposed results with the benchmarks as in Table 2 to highlight our hypothesis. For comparison, we include the solutions from the winning

Table 3 The evaluation in RSDO5 corpus given each domain as a test set in monolingual setting. Bold indicates the best result for each test set. The comparison between BIO and NOBI as well as the best model in F1-score are set in the same mechanism with Table 1

Valid set	Test set	BIO P	BIO R	BIO F1	NOBI P	NOBI R	NOBI F1
vet	ling	**69.6**	64.1	66.7	↓ 65.4	↑65.4	↓ 65.4
bim	ling	69.5	**73.7**	**71.5**	↓ **66.9**	↓ 69.5	↓ 68.2
kem	ling	66.2	72.4	69.2	↓ 64.9	↓ **72.3**	↓ **68.4**
ling	vet	71.1	66.7	68.8	↓ 66.6	↑68.5	↓ 67.5
kem	vet	**72.7**	65.6	**68.9**	↓ 66.9	↑ **69.7**	↓ **68.3**
bim	vet	69.3	**68.1**	68.7	↓ **67.6**	↓ 62.5	↓ 65.0
ling	kem	68.7	55.1	61.2	↓ 63.8	↑ **61.4**	↑ 62.6
bim	kem	**70.2**	**60.3**	**64.8**	↓ 66.1	↑ **61.4**	↓ 63.7
vet	kem	**70.2**	59.2	64.3	↓ **68.3**	↑ 60.6	↓ **64.2**
vet	bim	**63.5**	**66.8**	**65.1**	↓ **61.4**	↓ 61.3	↓ **61.3**
ling	bim	62.3	65.2	63.7	↓ 57.2	↓ 60.1	↓ 58.6
kem	bim	62.4	64.0	63.2	↓ 61.0	↓ **61.7**	↓ **61.3**
Avg.		68.0	65.1	66.3	↓ 64.7	↓ 64.5	↓ 64.5

Table 4 The evaluation in RSDO5 corpus given each domain as a test set in the multilingual setting. In this setting, in addition to Slovenian training data, the data from ACTER in en, fr, and nl is used, and ANN and NES training sets are compared

Valid. set	Test set	ANN BIO P	ANN BIO R	ANN BIO F1	ANN NOBI P	ANN NOBI R	ANN NOBI F1	NES BIO P	NES BIO R	NES BIO F1	NES NOBI P	NES NOBI R	NES NOBI F1
vet	ling	67.7	69.6	68.6	↓ **67.5**	↓ 62.7	↓ 65.0	67.2	**69.9**	**68.5**	↓ 64.2	↓ **67.3**	↓ **65.7**
bim	ling	**69.8**	66.2	67.9	↓ 64.6	↓ 68.1	↑ **66.3**	67.8	68.5	68.2	↓ **64.9**	↓ 64.8	↓ 64.8
kem	ling	66.5	**71.4**	**68.8**	↓ 59.6	↓ **71.0**	↓ 64.8	**67.9**	69.0	**68.5**	↓ 59.9	↓ 65.1	↓ 62.4
ling	vet	**71.0**	65.3	68.0	↓ 62.4	↑ **70.9**	↓ 66.4	69.2	67.4	68.3	↓ 61.8	↑ 70.8	↓ 66.0
kem	vet	69.8	**68.8**	**69.3**	↓ 68.0	↓ 68.5	↓ **68.2**	**70.5**	**67.8**	**69.1**	↓ **64.6**	↑ 70.6	↓ **67.5**
bim	vet	69.8	68.4	69.1	↓ **68.7**	↓ 67.1	↓ 67.9	69.3	64.7	66.9	↓ 63.0	↑ **72.8**	↓ **67.5**
ling	kem	68.3	59.3	63.5	↓ 66.0	↓ 52.9	↓ 58.7	67.5	54.6	60.4	↓ 62.8	↑ **60.8**	↑ 61.8
bim	kem	69.6	**61.2**	**65.1**	↓ **66.6**	↓ 55.5	↓ 60.5	**69.3**	52.7	59.9	↓ **65.5**	↑ **60.8**	↑ **63.1**
vet	kem	**69.9**	58.4	63.6	↓ 65.9	↓ **57.7**	↓ **61.5**	67.9	**59.2**	**63.3**	↓ 62.8	↑ **60.8**	↓ 61.8
vet	bim	61.2	**64.9**	**63.0**	↑ **62.9**	↓ 62.6	↓ 62.7	60.9	66.7	63.7	↓ 59.1	↓ 64.0	↓ 61.5
ling	bim	60.5	63.8	62.1	↓ 56.2	↓ 58.2	↓ 57.2	62.6	62.3	62.4	↓ 57.0	↑ 62.9	↓ 59.8
kem	bim	**65.7**	59.2	62.3	↓ 59.5	↑ **66.7**	↑ **62.9**	61.8	**67.1**	**64.3**	↓ 61.0	**67.1**	↓ **63.9**
Avg.		67.5	64.7	65.9	↓ 64.0	↓ 63.5	↓ 63.5	66.8	64.2	65.3	↓ 62.2	↑ 65.6	↓ 63.8

teams in the competition (TALN-LS2N (Hazem et al., 2020) won on the English and French test set, while NLPLab UQAM (Le & Sadat, 2021) won on the Dutch test set) and other methods (Rigouts et al., 2021; Lang et al., 2021) described in Sect. 2. Note that all the approaches from the related work are (1) cross-domain and (2) use the Heart failure domain as the test set, which shares the same mechanism with our approaches' validation.

Our proposed classifiers, trained using either BIO or NOBI annotation regimes, outperform previously described benchmark approaches, showcasing significant performance gains as measured by the F1. When comparing classifiers using BIO and NOBI annotation schemes, those utilizing BIO regimes demonstrate superior F1 on the English NES gold standard, which includes named entities. However, classifiers employing NOBI regimes exhibit noteworthy performance, surpassing all existing state-of-the-art (SOTA) models, including our BIO classifiers, across the languages present in both ANN and NES versions, with the exception of the aforementioned English NES corpus.

Table 5 Comparison between our performance and SOTA in RSDO5 dataset

Methods	Linguistics			Veterinary			Chemistry			Biomechanics		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SOTA Ljubešić et al. (2019)	52.2	25.4	34.1	66.9	19.3	29.9	47.8	31.4	37.8	53.8	24.8	33.9
Mono BIO	69.5	73.7	71.5	72.7	65.6	68.9	70.1	60.3	64.8	63.5	66.8	65.1
Multi BIO	66.5	71.5	68.8	69.8	**68.8**	**69.3**	69.6	**61.2**	**65.1**	61.8	**67.1**	64.3
Mono NOBI	64.9	72.3	68.4	66.9	69.7	68.3	68.3	60.6	64.2	61.4	61.3	61.3
Multi NOBI	64.6	68.1	66.3	68.0	68.5	68.2	65.5	60.8	63.1	61.0	67.1	63.9

Furthermore, we conduct a multilingual evaluation to examine the impact of adding additional languages to the training set. In contrast to the findings of Lang et al. (2021), we observe that incorporating other languages generally leads to only marginal improvements in model performance.

4.2 Evaluation on the RSDO5 test set

We also apply monolingual and multilingual cross-domain approaches to the Slovenian RSDO5 dataset. The results grouped by the test domain using BIO and NOBI annotation regimes are presented in Tables 3 and 4, respectively. For each annotation regime, we evaluate monolingual and multilingual settings where ANN and NES versions are added to the training set of the RSDO5 corpus.

The monolingual approach, where we use two domains from the RSDO5 corpus for training, validate on the third domain, and test on the last domain, proves to have relatively consistent performance across all the combinations in both annotation regimes. For both regimes, we achieve a Precision of more than 61%, Recall of no less than 55%, and F1 above 57%. Furthermore, they perform slightly better in the Linguistics and Veterinary domains than in Biomechanics and Chemistry. The difference in the number of terms and length of terms per domain pointed out in Sect. 3.1 might be one of the factors that contribute to this behavior. Moreover, a significant performance boost can be observed for the Veterinary domain when the model is trained in the Biomechanics and Linguistics domains and for the Linguistics domain if the Veterinary domain is included in the training set for the model in both annotation regimes. Between these two settings, the classifier with BIO regime gained a performance of up to 68.9% in the F1 for the Linguistics test set, which surpasses other domains in the same regimes as well as outperforms all the cases in the monolingual classifier of the NOBI regime.

We also explore the performance of multilingual approaches on the RSDO5 test sets. We train the model using the ANN and NES labels from all domains of the ACTER dataset and on two domains from the RSDO5 dataset, validate on the third RSDO5 domain, and test on the last domain. Table 3 and 4 present the comparative performance of the multilingual and the monolingual approaches. However, from the results, there exists a discrepancy in the performance-boosting efficiency among the different combinations of training, validation, and test sets. This raises a hypothesis of the domain sensitivity in transfer learning for ATE tasks. Thus, a careful choice of the domains in the training set is undoubtedly necessary for boosting the classifier's performance.

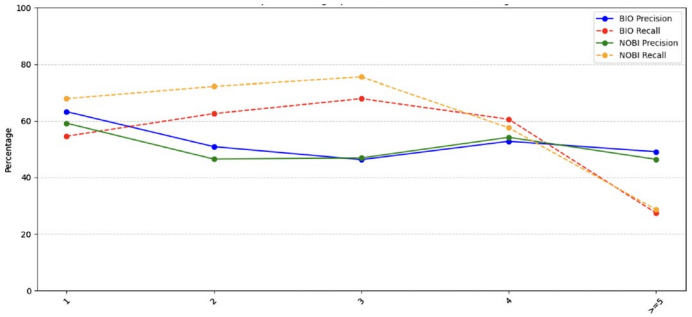

Fig. 5 Performance in P and R per term length per domain in English ACTER test set

Besides, we compare two different annotation regimes by evaluating the performance of classifiers using different training, validation, and testing combinations for each regime. Despite the consistency in the predictive power of monolingual and multilingual settings, the classifiers with NOBI annotation presented a worse performance in the Slovenian RSDO5 corpus compared to the BIO regime. This is due to the fact that the proportion of nested terms in RSDO5 is too small for the classifier to learn nested terms properly, which are visualized in the proportion of unique nested terms and terms nested in other nested terms from Figs. 16 to 18.

In Table 5, we present the results from the related work for the RSDO5 dataset compared to our proposed monolingual and multilingual approaches. The result from Ljubešić et al. (2019)'s method, which has been re-implemented using the same RSDO5 corpus as our studies, is taken from Tran et al. (2022b). In general, our approach outperforms Ljubešić et al. (2019)'s one by a large margin on all domains and according to all evaluation metrics, especially when it comes to Recall. We achieve results roughly twice as high as Ljubešić et al. (2019)'s approach in F1-score for all test domains regarding both monolingual and multilingual learning. One should note that the method (Ljubešić et al., 2019) was primarily meant for extracting terms from Ph.D. theses, i.e., documents significantly longer than those available in our training data, which explains the low Recall of that approach. However, this result clearly identifies a significant strength of the sequence-labeling approach - it does not rely on the frequency of term occurrences, which makes the approach more robust as shown in this comparison. In our case, we show that the multilingual experiments do in several cases improve our monolingual results (Tran et al., 2022b), but not systematically.

5 Error analysis

In order to determine whether the term length affects the models' performance, we calculate Precision and Recall for terms of length k = {1,2,3,4, ≥5} when predicted by our classifiers on the test set. The number of predicted candidate terms (Preds), number of ground truth terms (GTs), number of correct predictions (TPs), Precision (P), and Recall (R) regarding different term lengths k and test domains in ACTER and

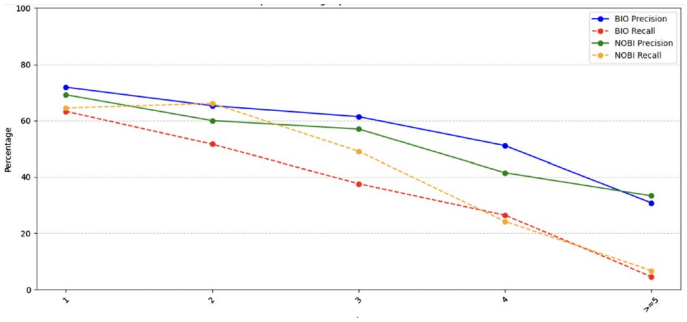

Fig. 6 Performance in P and R per term length per domain in French ACTER test set

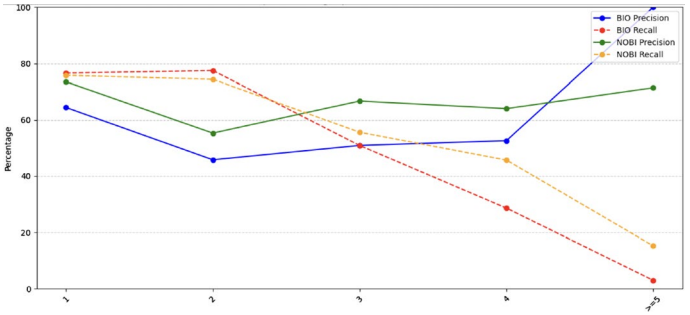

Fig. 7 Performance in P and R per term length per domain in Dutch ACTER test set

RSDO5 corpora are presented in Table 9 and 10 (in Appendix) and Precision (P) and Recall (R) of each scenario are visualized below.

5.1 The ACTER dataset

The results for ACTER's dataset (Table 9) were obtained by employing the best performing model for a specific language in terms of F1 on the Heart failure test set for the most cases (which is the combination of English, French, and Slovenian as the training set).

As demonstrated in Fig. 5, 6, and 7, when using the BIO scheme, the best model proved to be good at predicting terms containing up to four words for English and Dutch and up to three words for French texts in ACTER corpora. A strong correspondence between the F1 and the number of predicted candidate terms has been found where the number of predicted candidate terms likely corresponds to the situation in the training data (see Table 9 in Appendix).

The best models trained using the NOBI annotation scheme demonstrated the same behavior as the one trained using the BIO annotation regime. They performed well at predicting terms containing up to four words for English and Dutch and up to three words for French texts in ACTER corpora. While our expectation was that the NOBI annotation scheme should benefit the model's ability to predict short one-word nested terms, the classifiers trained using NOBI annotations show better performance than those using the BIO regime on multi-word terms as well, as long as nested terms take a proper proportion as

Table 6 A comparison of the performance between the BIO and NOBI schemes on the entire dataset, single-word terms (SWU), and multi-word terms (MWU)

	BIO			NOBI		
	P	R	F1	P	R	F1
All terms	58.1	48.1	52.6	57.5	48.6	52.7
SWU	65.0	45.9	53.8	61.6	51.5	56.1
MWU	53.8	50.0	51.8	54.2	46.3	49.9

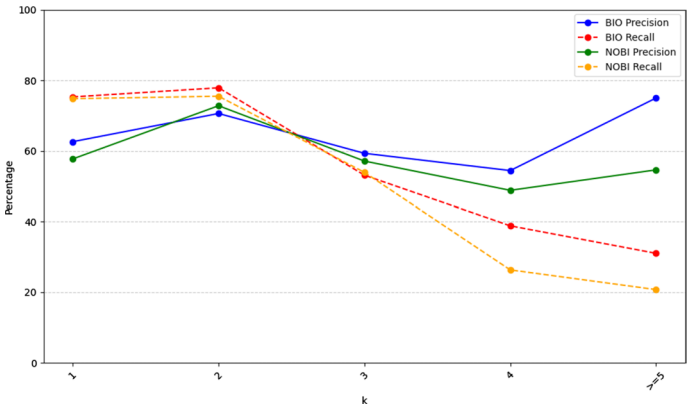

Fig. 8 Performance in P and R per term length per domain in RSDO Linguistics test set

in ACTER corpora. The Recall therefore generally improves for terms of all lengths, even for terms containing 5 words or more. There seems to be some signal in the occurrence of nested terms inside multi-word terms, which leads the model to better identify longer terms as well. Our current hypothesis is that this effect is a combination of (1) the improvement of single-word term identification by having a larger training set available (both nested and independent single-word terms) and (2) nested terms being some sort of anchor exploited by the model to easier identify multi-word terms around that nested terms. Further experiments and analyses should be conducted to fully understand this phenomenon.

Furthermore, a trend that is noticeable across the majority of scenarios is that the NOBI regime reduces the Precision compared to the BIO regime. This seems to be related to the number of terms predicted where we can observe that Precision often drops where the number of predicted terms is higher, i.e., the BIO regime on the English dataset predicts 1,009 single-word terms with a Precision of 63.3 % and the NOBI regime predicts 1,341 terms with a Precision of 59.2%. In a similar but reversed trend, the Dutch NOBI regime produced 1,738 terms with a Precision of 73.5% whereas the BIO regime produced 2005 terms with a Precision of 64.4% (see Table 9 for the statistics).

We performed an additional detailed comparison of the BIO and NOBI monolingual results on the English dataset (i.e., the results from the first line in Table 1) in Table 6. The NOBI scheme produces a marginal improvement in terms of F1 and Recall but has slightly lower Precision. Overall, the algorithm predicted 1,956 candidates when using the BIO scheme and 1,996 when using the NOBI scheme. Out of these, the BIO scheme resulted in 751 single-word terms (SWU) and 1205 multi-word terms (MWU), while the NOBI scheme produced 889 single-word terms and 1,107 multi-word terms. Looking at the performance in Table 6, NOBI results in a better Recall of single-word terms (51.5 vs. 45.9),

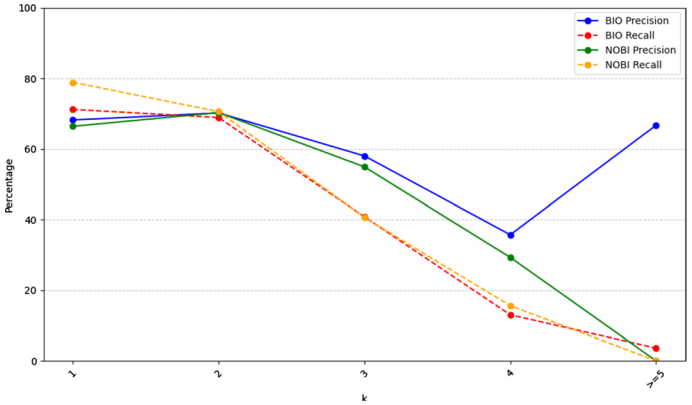

Fig. 9 Performance in P and R per term length per domain in RSDO Veterinary test set

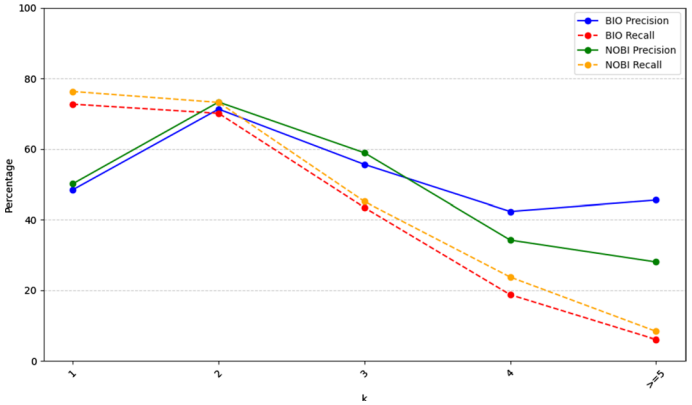

Fig. 10 Performance in P and R per term length per domain in RSDO Biomechanics test set

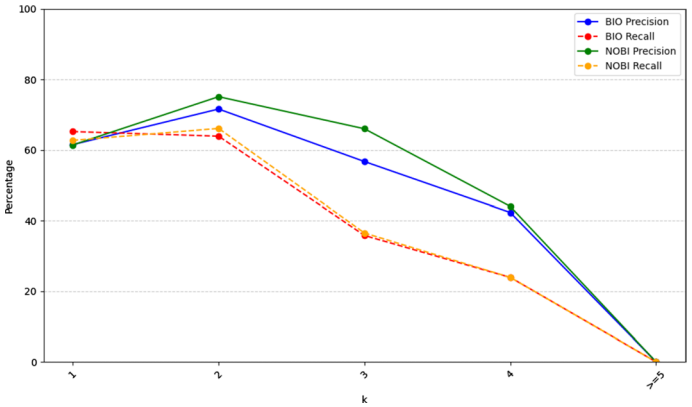

Fig. 11 Performance in P and R per term length per domain in RSDO Chemistry test set

which leads to an overall improvement of the F1 (52.7 vs. 52.6). It does not improve the Precision of SWU terms, but does, perhaps surprisingly, deliver higher Precision on MWU terms, which could be due to the fact that the NOBI regime prefers single-word terms (due to their higher proportion in the training set) which results in a smaller number of higher quality MWU terms being predicted.

5.2 The RSDO5 dataset

The results for the RSDO5 dataset (Table 10 in Appendix and from Figs. 8, 9, 10 and 11) were obtained by employing the best-performing model in the F1 for each specific test domain for both annotation regimes, which are (1) training on Veterinary and Chemistry, validation on Biomechanics, and testing on Linguistics domain; (2) training on Linguistics and Biomechanics, validation on Chemistry, and testing on Veterinary domain; (3) training on Linguistics and Veterinary, validation on Biomechanics, and testing on Chemistry domain; (4) training on Linguistics and Chemistry, validation on Veterinary, and testing on Biomechanics.

These results are similar to ACTER corpora, showing that the models are good at predicting short terms containing up to three words for all four domains of the Slovenian corpus. The best model applied to the Linguistics test domain also shows relatively good performance when it comes to the prediction of longer terms, achieving 75.0% Precision and a decent 31.0% Recall for terms with at least five words. Despite the relatively high Precision for prediction of long terms in the Veterinary and Biomechanics test domains, the Recall is pretty low, most likely due to the small amount of longer terms in the dataset on which the models are trained. When predicting the Chemistry domain, there are no correct predictions of more than five-word terms.

The NOBI regime often results in a lower Precision compared to the BIO one. Similar to our findings on the ACTER dataset, this seems to be related to the number of terms being predicted. In general, the higher the number of predictions, the lower the Precision (if the number of predicted terms is high enough — this trend is less noticeable for longer terms of which there are few in the corpus). There are some exceptions, like the Chemistry domain, where the NOBI regime results in 909 predicted single-word terms with a Precision of 61.4% compared to 943 terms with a Precision of 61.5% for the BIO regime, and the Veterinary domain where the NOBI regime predicted 2,111 two-word terms (k=2) with a Precision of 70.3% while the BIO regime predicted 2062 terms with a Precision of 70.2%.

As mentioned above, as well as in previous work (Tran et al., 2022b) for the BIO regime, since the corpus contains nested terms, the very common mistake the both BIO and NOBI models make is to incorrectly predict a shorter term nested in the correct term of the gold standard. Vice versa, the model sometimes generates incorrect predictions containing the correct nested terms. However, the NOBI annotation proves to partially reduce the effect of these two mentioned error patterns and improves the general Recall in comparison to the benchmark BIO scheme.

6 Conclusion

In summary, we demonstrated the possibilities of cross- and multilingual learning compared to the monolingual setting in boosting the predictive performance of the cross-domain sequence-labeling term extraction via experiments conducted on multi-domain

corpora, namely the ACTER and RSDO5 datasets. In addition, we presented the positive impact of cross- and multilingual models on the ACTER corpora only, and by further adding the texts from the Slovenian RSDO5 corpus in the training set. Furthermore, we examined the cross-lingual effect of rich-resourced training language on less-resourced testing ones such as Slovenian. Last but not least, we proposed a new NOBI annotation regime, that boosted the predictive power of classifiers in comparison to the classical BIO mechanism, as shown in the ACTER corpus, in which the number of nested terms is significant enough. The improvements through the NOBI annotation regime are visible even in multi-word term identification, quite likely by improving single-word term extraction and exploiting single-word terms as anchors to correctly identify multi-word terms. The results demonstrated the potential of the new annotation scheme to enhance the nested term extraction and a promising impact of cross- and multilingual cross-domain learning when transferring from rich- to less-resourced languages.

In future work, we will test the potential of our proposed NOBI mechanism in similar sequence-labeling extraction tasks in other domains (e.g., Named Entity Recognition). In addition, we plan to investigate the integration of active learning into our current approach to improve the output of the automated method by dynamical adaptation after human feedback.

Appendix

Data analysis

Figure 12 presents the structures of two datasets that we used for our work, including ACTER corpora and RSDO5 corpus. Note that in ACTER datasets, two versions of the gold standard were proposed: (1) ANN version covering only terms; and (2) NES version including both terms and named entities.

Figure 13 illustrates an example of the key difference between the ACTER's ANN and NES versions of gold standards. Given the sentence "*...This study uses the Medicare Patient Safety Monitoring System...*", the gold standard of the ANN version consists of only the term "*Patient*" as the only term was annotated as the ground truth. On the other hand, the NES version's gold standard includes the Named Entity (NE) "*Medicare Patient Safety Monitoring System*" as both domain-specific terms and NEs were annotated in the ground truth.

Figure 14 summarizes the number of unique terms (e.g., the term counts excluding the duplication) for each domain in both ACTER and the Slovenian RSDO5 corpora. It contains statistics for both ANN (annotating only terms) and NES (annotating both terms and named entities as the ground truth terms) versions in the ACTER set. This supports the statements in Subsection 3.1 and 3.2.

Table 7 indicates the proportion of the nested terms with different word length k where $k = \{1, 2, 3, 4, \geq 5\}$ for each domain and language of both corpora, which also supports to the statements in Subsection 3.3. The last column on the right calculates the percentage of single-word nested terms in total nested terms in the first level. On average, the amount of single-word nested terms accounts for 78.06% above all the nested terms on the first levels in the corpora. That is why we do not consider either multi-word nested terms or terms nested in other nested terms - so-called nested terms on the second or higher levels and we label single-word nested terms on the first level.

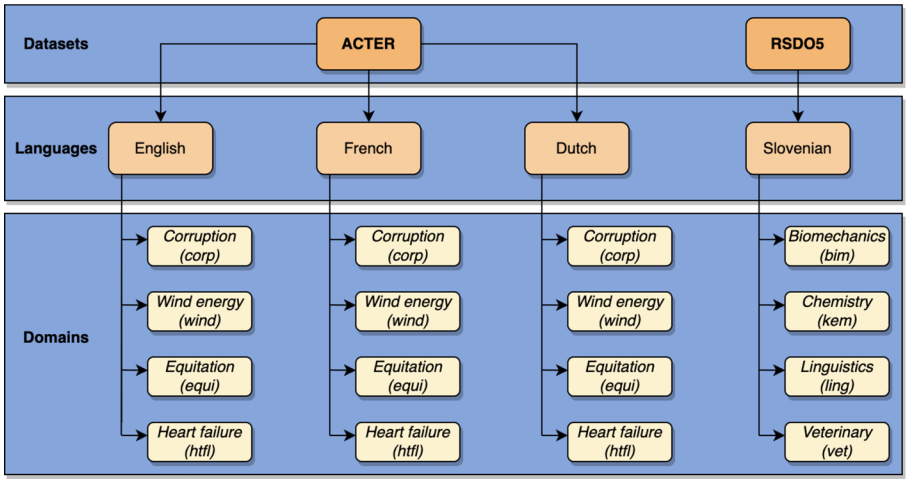

Fig. 12 The structure of RSDO5 and ACTER regarding languages and domains

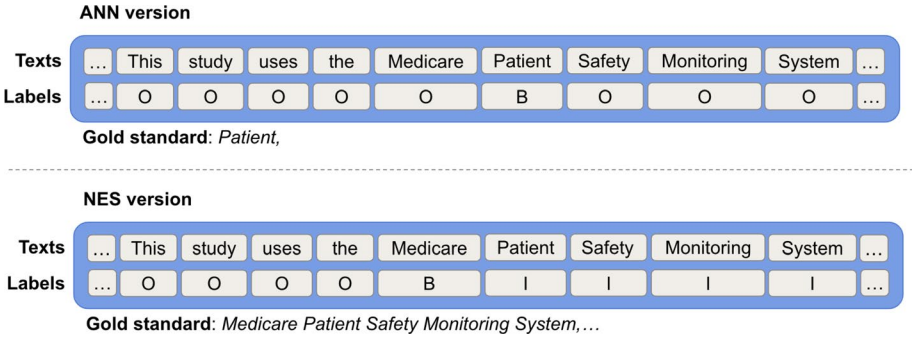

Fig. 13 An example of ACTER's ANN and NES versions were annotated in the BIO regime

Regarding ACTER corpora, Figs. 15 and 16 present the term density and the proportion of unique nested terms founded in texts extracted from the ACTER corpora for each domain and language, respectively. As can be seen from both figures, a notable disparity in data volume and term distribution is observed, particularly between the Heart failure domain and the other three domains, with the former containing a more significant number of unique terms. Further comprehensive information on the ACTER dataset can be found in the TermEval competition by Rigouts et al. (2020a).

Similarly, Figs. 17 and 18 present the term density and the unique term proportion in texts captured from the RSDO5 corpus for each domain, respectively. As can be seen, the documents from the Linguistics and Veterinary domains contain more terms than Biomechanics and Chemistry. Most terms are made of up to three words and only a few terms are longer than seven words. For example, an observation of the long multi-word term found in the corpus would be *"stojo po obračanju v nasprotni smeri urinega kazalca"* (stand after turning counterclockwise) in Biomechanics.

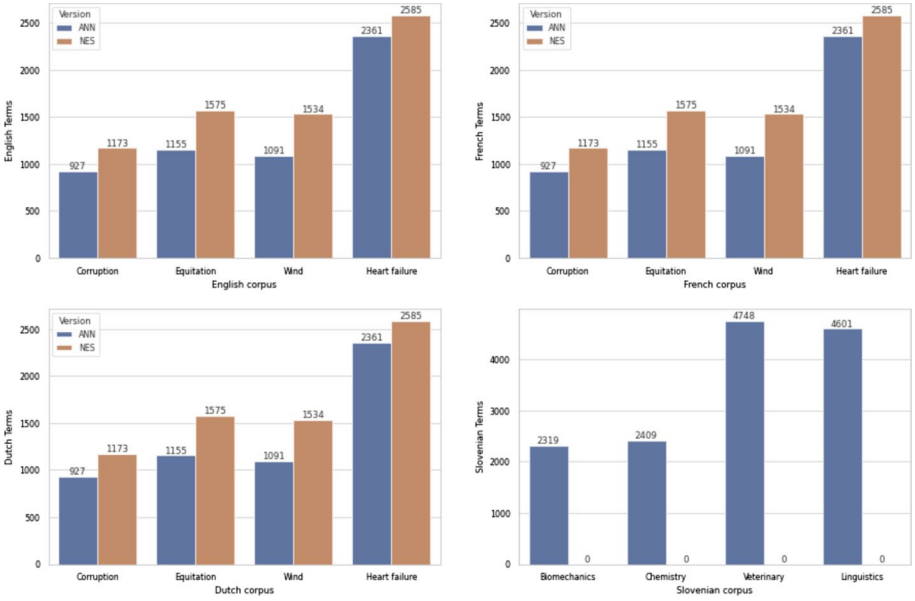

Fig. 14 The number of unique terms in ACTER and RSDO5 corpora

Table 7 The proportion of unique nested terms of different word lengths in each domain and language of ACTER and RSDO5 corpora

Languages	Domains	k = 1	k = 2	k = 3	k = 4	k ≥ 5	% (k = 1)
ACTER							
en	corp	246	89	11	1	1	70.69
	equi	469	87	5	1	0	77.90
	wind	282	171	36	4	0	83.51
	htfl	580	183	55	20	6	83.45
fr	corp	289	59	19	2	2	87.60
	equi	339	32	13	3	0	86.97
	wind	192	38	24	6	1	57.20
	htfl	620	99	30	8	9	73.56
nl	corp	309	46	12	2	1	84.90
	equi	414	44	12	6	0	68.72
	wind	253	36	4	4	1	80.94
	htfl	574	46	4	4	0	91.40
RSDO5							
sl	ling	737	177	8	0	0	79.93
	vet	835	199	13	5	1	79.30
	kem	388	126	7	2	1	74.05
	bim	349	111	17	16	14	68.84
Average							78.06

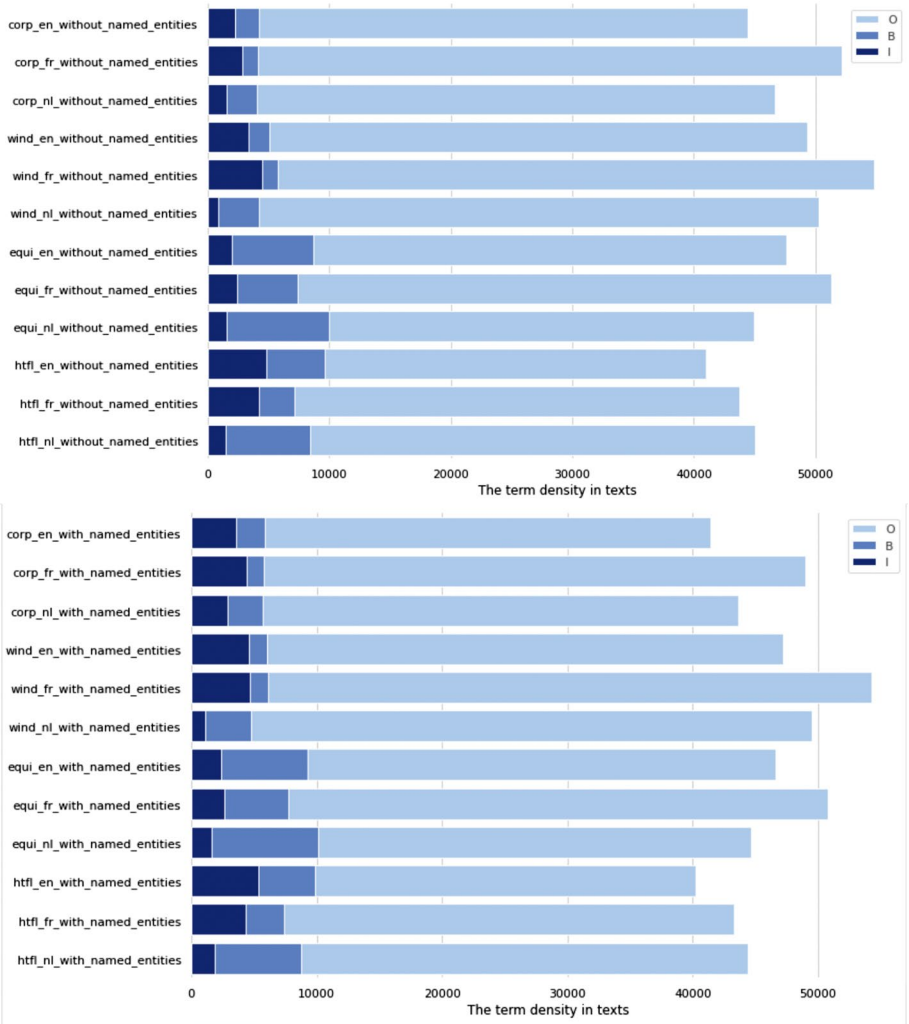

Fig. 15 The term density in BIO regime in the ACTER corpora

Annotation regimes

Besides the popular BIO regime, IOBES and BILOU are two different annotation schemes commonly used in Natural Language Processing (NLP) tasks. These schemes are used to represent and label entities within a sequence of words or tokens in a text. IOBES stands for tokens [**I**]nside an entity; [**O**]utside an entity (i.e., not part of any entity); [**B**]eginning token of an entity; [**E**]nd token of an entity; [**S**]ingle token that forms a whole entity by itself. Compared to the BIO scheme, the IOBES scheme is an extension of the BIO scheme with the "E" and "S" tags added to represent entities that end at a token or consist of a single token. Meanwhile, BILOU represents tokens [**B**]eginning token of an entity; [**I**]nside an entity; [**L**]ast token of an entity; [**O**]utside an entity; and [**U**]nit token that forms a whole entity by itself. Sharing the same "B", "I", and "O", the BILOU scheme is an extension

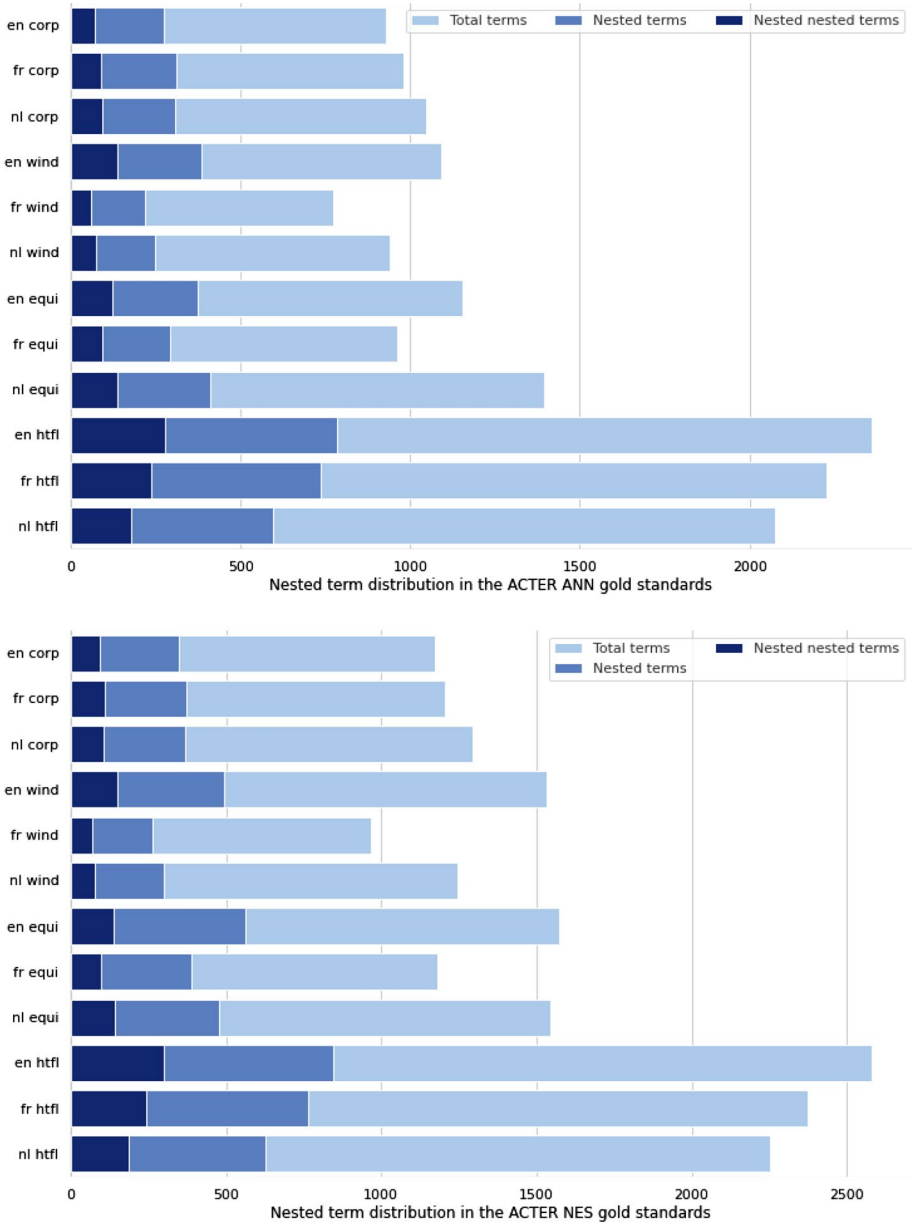

Fig. 16 The proportion of unique nested terms in the ACTER gold standards

of the IOBES scheme, but it offers a more compact representation of entities that consist of multiple tokens. We reported the performance of our XLMR classifier fine-tuning on ACTER English sets in BIO, NOBI, BIOES, and BILOU with ANN gold standard as demonstrated in Table 8.

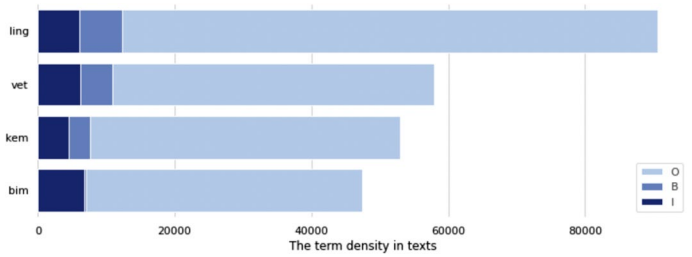

Fig. 17 The term density in BIO regime per domain of the RSDO5 corpus

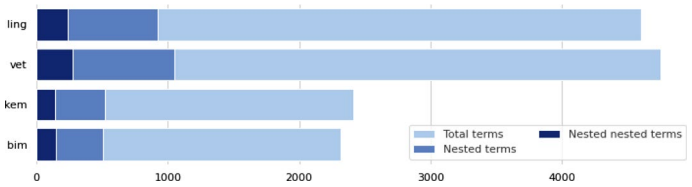

Fig. 18 The proportion of unique nested terms in the RSDO5 gold standards

The results demonstrate the superiority of our novel annotation regimes in comparison with other related schemes. In fact, both IOBES and BILOU are widely used to label entities and are often converted into simpler BIO formats during training or evaluation. These annotation schemes help models understand the boundaries and types of entities present in a text, enabling them to learn to recognize and extract them effectively. However, the standard IOBES and BILOU annotation schemes do not well support nested entities but were used as a foundation (similar to BIO) to improvise on the nested terms. Both IOBES and BILOU are designed to represent terms or entities in a flat manner, where each token in the text is associated with only one entity tag. In a nested entity scenario, we would have ones that are hierarchically structured, with one entity/term fully or partially contained within another entity/term. To represent nested entities, we propose more custom annotation schemes, namely NOBI, and a simple designed scheme to handle the single-word nested structures appropriately.

Monolingual vs. multilingual pre-trained models

We evaluated the performance using monolingual language models, including XLNet[5] (English), CamemBERT[6] (French), and DutchBERT[7] (Dutch) compared against a multilingual model, XLMR[8], which was pre-trained on over 100 different languages and fine-tuned for the downstream ATE task, as visualized from Figs. 19, 20, 21. The selection of the monolingual models is based on their superior performance in the empirical evaluation of various monolingual and multilingual Transformer-based models on monolingual sequence-labeling cross-domain term extraction (Tran et al., 2022c).

[5] xlnet-base-cased (https://huggingface.co/xlnet-base-cased).
[6] camembert-base (https://huggingface.co/camembert-base).
[7] GroNLP/bert-base-dutch-cased (https://huggingface.co/GroNLP/bert-base-dutch-cased).
[8] xlm-roberta-base (https://huggingface.co/xlm-roberta-base).

Table 8 Evaluation of XLMR classifier fine-tuning on ACTER English sets in BIO, NOBI, BIOES, and BILOU with NES gold standard

Models	P	R	F1
BIO	62.1	52.1	56.7
BIOES	62.6	51.9	56.7
BILOU	61.8	52.6	56.8
NOBI	58.6	55.2	56.9

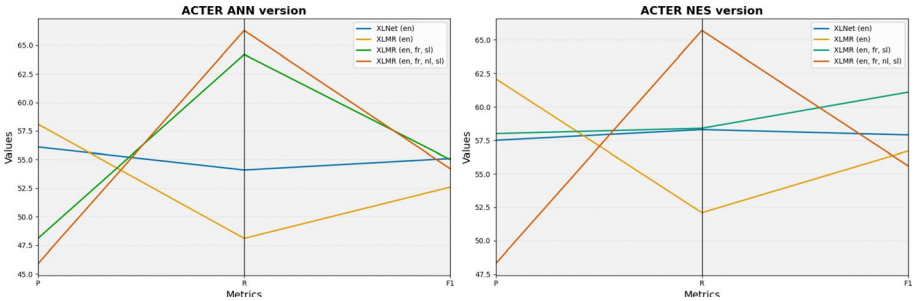

Fig. 19 Performance of monolingual pre-trained classifier finetuned on English test language vs. multilingual one finetuned on the test language and multiple languages in ACTER

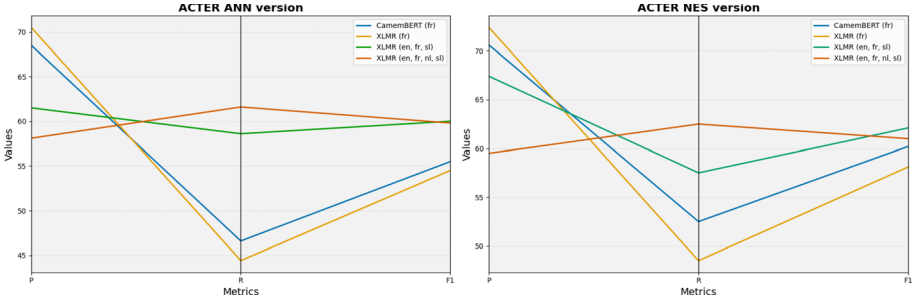

Fig. 20 Performance of monolingual pre-trained classifier finetuned on French test language vs. multilingual one finetuned on the test language and multiple languages in ACTER

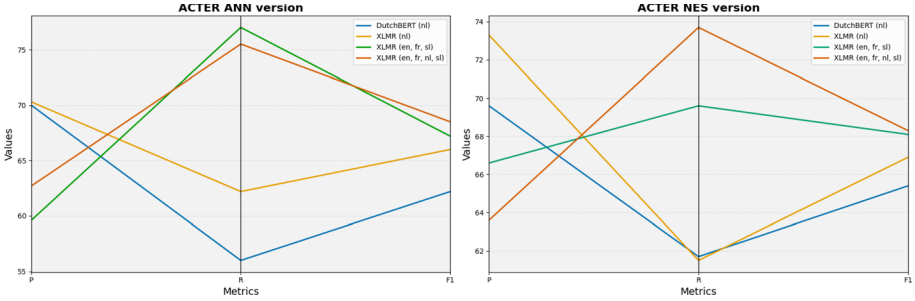

Fig. 21 Performance of monolingual pre-trained classifier finetuned on Dutch test language vs. multilingual one finetuned on the test language and multiple languages in ACTER

Table 9 Performance per term length per domain in ACTER dataset

k	English					French					Dutch				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
BIO regime															
1	1,009	1,170	639	63.3	54.6	1,153	1,309	829	71.9	63.3	2,005	1,687	1,292	64.4	76.6
2	985	801	501	50.9	62.6	490	620	320	65.3	51.6	661	391	303	45.8	77.5
3	553	377	256	46.3	67.9	163	266	100	61.4	37.6	108	108	55	50.9	50.9
4	163	142	86	52.8	60.6	47	91	24	51.1	26.4	19	35	10	52.6	28.6
≥5	53	95	26	49.1	27.4	13	88	4	30.8	4.6	1	33	1	100.0	3.0
NOBI regime															
1	1,341	1,170	794	59.2	67.9	1,219	1309	844	69.2	64.5	1,738	1,687	1,278	73.5	75.8
2	1,242	801	578	46.5	72.2	683	620	410	60.0	66.1	526	391	291	55.3	74.4
3	606	377	284	46.9	75.5	228	266	130	57.0	49.1	90	108	60	66.7	55.6
4	153	142	83	54.2	57.6	53	91	22	41.5	24.2	25	35	16	64.0	45.7
≥5	56	95	26	46.4	28.6	18	88	6	33.3	6.7	7	33	5	71.4	15.2

The results using the monolingual models exhibit slightly higher performance in the specific language they were pre-trained on. However, when applied in a cross-lingual context (e.g., fine-tuning XLNet on an English corpus and predicting on a French test set), the performance is significantly diminished when compared to the multilingual pre-trained model (e.g., XLMR). While the difference between the language-specific and multilingual models is small, the multilingual models, trained with XLMR on the datasets of multiple and all languages, for the most part, outperform the monolingual models by a small margin. As a result, in order to accommodate and support multiple languages simultaneously, we opt to utilize XLMR as the benchmark model for all four languages in ACTER and RSDO5 corpora to validate our hypothesis in this study.

Error analysis

We calculate Precision and Recall for terms of length k = {1,2,3,4, ≥5} when our classifiers predict on the test set. The number of predicted candidate terms (Preds), number of ground truth terms (GTs), number of correct predictions (TPs), Precision (P), and Recall (R) regarding different term lengths k and test domains in ACTER and RSDO5 corpora are presented in Table 9 and 10.

These Tables provide detailed support for the explanation of the classifier's behavior toward each dataset in terms of term length.

Table 10 Performance per term length per domain in RSDO corpus

NOBI regime										
k	Linguistics					Veterinary				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	2,078	1,728	1,300	62.6	75.3	2,159	2,067	1,472	68.2	71.2
2	2,631	2,404	1,858	70.6	77.9	2,062	2,103	1,448	70.2	68.9
3	322	360	7,191	59.3	53.1	314	446	182	58.0	40.8
4	57	80	31	54.4	38.8	28	77	10	35.7	13.0
≥5	12	29	79	75.0	31.0	3	55	2	66.7	3.6
k	Chemistry					Biomechanics				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	943	890	580	61.5	65.2	1,079	718	22	48.4	72.7
2	1,073	1,202	768	71.6	63.9	1,153	1,172	822	71.3	70.1
3	164	260	93	56.7	35.8	223	286	124	55.6	43.4
4	26	46	11	42.3	23.9	26	59	11	42.3	18.7
≥5	3	11	0	0.0	0.0	11	84	5	45.5	6.0
NOBI regime										
k	Linguistics					Veterinary				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	2241	1728	1293	57.7	74.8	2456	2067	1630	66.4	78.9
2	2491	2404	1814	72.8	75.5	2111	2103	1484	70.3	70.6
3	340	360	194	57.1	53.9	330	446	181	54.9	40.6
4	43	80	21	48.8	26.3	41	77	12	29.3	15.6
≥5	11	29	6	54.6	20.7	5	55	0	0.0	0.0
k	Chemistry					Biomechanics				
	Preds	GTs	TPs	P	R	Preds	GTs	TPs	P	R
1	909	890	558	61.4	62.7	1094	718	548	50.1	76.3
2	1058	1202	795	75.1	66.1	1171	1172	858	73.3	73.2
3	144	260	95	66.0	36.5	219	286	129	58.9	45.1
4	25	46	11	44.0	23.9	41	59	14	34.2	23.7
≥5	0	11	0	0	0.0	25	84	7	28.0	8.3

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Hanh Thi Hong Tran and Matej Martinc. The baseline method was proposed by Nikola Ljubesic. The first draft of the manuscript was written by Hanh Thi Hong Tran and all authors commented on previous versions of the manuscript. Andraz Repar proposed the analysis with respect to the difference between NOBI and BIO behavior in predicting the candidate terms. All authors read and approved the final manuscript.

Funding The work was partially supported by the Slovenian Research Agency (ARIS) via the core research programs Knowledge Technologies (P2-0103) and Language resources and technologies for Slovene (P6-0411), project Formant Combinatorics in Slovenian (J6-3131), as well as by the Ministry of Culture of

the Republic of Slovenia through the project Development of Slovene in Digital Environment (RSDO). The first author was partly funded by Region Nouvelle Aquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France. The work was also supported by the project Cross-lingual and cross-domain methods for Terminology Extraction and Alignment, a bilateral project funded by the program PROTEUS under the grant number BI-FR/23-24-PROTEUS006.

Data availability and materials The original datasets are collected from two sources: ACTER version 1.5 (https://github.com/AylaRT/ACTER) and RSDO version 1.1 (https://www.clarin.si/repository/xmlui/handle/11356/1470). The newly annotated dataset is publicly available at https://github.com/honghanhh/nobi_annotation_regime.git.

Code availability Our code is publicly available at https://github.com/honghanhh/ate_nobi.git.

Declarations

Conflict of interest Not applicable.

Ethical approval Not applicable.

Consent to participate All the authors consent to participate.

Consent for publication All the authors consent for publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (pp. 54–59).

Amjadian, E., Inkpen, D., Paribakht, T., & Faez, F. (2016). Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)* (pp. 2–11).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL*.

Daille, B., Gaussier, É., & Langé, J. M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Damerau, F. J. (1990). Evaluating computer-generated domain-oriented vocabularies. *Information Processing and Management, 26*(6), 791–801.

ElKishky, A., Song, Y., Wangx, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment, 8*(3), 305–316.

Erjavec, T., Fišer, D., & Ljubešić, N. (2021). The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation, 55*(2), 551–583.

Fišer, D., Suchomel, V., & Jakubícek, M. (2016). Terminology extraction for academic slovene using sketch engine. In *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016* (pp. 135–141).

Frantzi, K.T., Ananiadou, S., & Tsujii, J. (1998). The c-value/nc-value method of automatic recognition for multi-word terms. In *International conference on theory and practice of digital libraries* (pp. 585–604). Springer.

Gao, Y., & Yuan, Y. (2019). Feature-less End-to-end Nested Term extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 607–616). Springer.

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2020). TermEval 2020: TALN-LS2N System for Automatic Term Extraction. In *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 95–100).

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2022). Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 648–662).

Jemec Tomazin, M., Trojar, M., Atelšek, S., Fajfar, T., Erjavec, T., & Žagar Karer, M. (2021). Corpus of term-annotated texts RSDO5 1.1. URL http://hdl.handle.net/11356/1470. Slovenian language resource repository CLARIN.SI

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1*(1), 9–27.

Kessler, R., Béchet, N., & Berio, G. (2019). Extraction of terminology in the field of construction. In *2019 First International Conference on Digital Data Processing (DDP)* (pp. 22–26). IEEE.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., & Stüker, S. (2018). Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In *INTERSPEECH* (pp. 2072–2076).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 260–270).

Lang, C., Wachowiak, L., Heinisch, B., & Gromann, D. (2021). Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 3607–3620).

Le, N. T., & Sadat, F. (2021). Multilingual automatic term extraction in low-resource domains. In *The International FLAIRS Conference Proceedings*, vol. 34.

Le Serrec, A., L'Homme, M. C., Drouin, P., & Kraif, O. (2010). Automating the compilation of specialized dictionaries: Use and analysis of term extraction and lexical alignment. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication, 16*(1), 77–106.

Lester, B. (2020). iobes: A library for span-level processing. arXiv preprint arXiv:2010.04373.

Lingpeng, Y., Donghong, J., Guodong, Z., & Yu, N. (2005). Improving retrieval effectiveness by using key terms in top retrieved documents. In *European Conference on Information Retrieval* (pp. 169–184). Springer.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics, 8*, 726–742.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). Kas-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. In *International Conference on Text, Speech, and Dialogue* (pp. 115–126). Springer.

Marciniak, M., & Mykowiecka, A. (2015). Nested term recognition driven by word connection strength. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication, 21*(2), 180–204.

Martinc, M., Škrlj, B., & Pollak, S. (2021). Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering* (pp. 1–40). https://doi.org/10.1017/S1351324921000127

Nugumanova, A., Akhmed-Zaki, D., Mansurova, M., Baiburin, Y., & Maulit, A. (2022). NMF-based approach to automatic term extraction. *Expert Systems with Applications, 199*, 117179.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostaja, T., Vintar, Š., & Fišer, D. (2019). Extracting data from comparable corpora. In *Using Comparable Corpora for Under-Resourced Areas of Machine Translation* (pp. 89–139). Springer.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. *Natural language processing using very large corpora* (pp. 157–176).

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)* (pp. 147–155).

Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., & Pollak, S. (2019). TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment. *International Journal of Theoretical and Applied Issues in Specialized Communication, 25*(1), 93–120.

Rigouts Terryn, A., Hoste, V., Drouin, P., & Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)* (pp. 85–94). European Language Resources Association (ELRA).

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2020). In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation, 54*(2), 385–418.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2021). HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology*.

Tran, H. T. H., Doucet, A., Sidere, N., Moreno, J. G., & Pollak, S. (2021). Named entity recognition architecture combining contextual and global. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021*, Virtual Event, December 1–3, 2021, Proceedings, p. 264. Springer Nature.

Tran, H. T. H., Martinc, M., Doucet, A., & Pollak, S. (2022). Can cross-domain term extraction benefit from cross-lingual transfer? In *International Conference on Discovery Science* (pp. 363–378). Springer.

Tran, H. T. H., Martinc, M., Doucet, A., & Pollak, S. (2022). A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction. In *Slovenian Conference on Language Technologies and Digital Humanities*.

Tran, H. T. H., Martinc, M., Pelicon, A., Doucet, A., & Pollak, S. (2022). Ensembling transformers for cross-domain automatic term extraction. In *International Conference on Asian Digital Libraries* (pp. 90–100). Springer.

Vintar, Š. (2004). Comparative evaluation of c-value in the treatment of nested terms. In *Workshop Description* (pp. 54–57).

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

Wolf, P., Bernardi, U., Federmann, C., & Hunsicker, S. (2011). From statistical term extraction to hybrid machine translation. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Zhang, Z., Gao, J., & Ciravegna, F. (2018). Semre-rank: Improving automatic term extraction by incorporating semantic relatedness with personalised pagerank. *ACM Transactions on Knowledge Discovery from Data (TKDD), 12*(5), 1–41.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.