Corpora for Bilingual Terminology Extraction in Cybersecurity Domain

Andrius Utka
Vytautas Magnus University, CCL
Kaunas, Lithuania
andrius.utka@vdu.lt

Sigita Rackevičienė
Mykolas Romeris University
Vilnius, Lithuania
sigita.rackeviciene@mruni.eu

Liudmila Mockienė
Mykolas Romeris University
Vilnius, Lithuania
liudmila@mruni.eu

Aivaras Rokas
Vytautas Magnus University, CCL
Kaunas, Lithuania
aivaras.rokas@vdu.lt

Marius Laurinaitis
Mykolas Romeris University
Vilnius, Lithuania
laurinaitis@mruni.eu

Agnė Bielinskienė
Vytautas Magnus University
Kaunas, Lithuanian
agne.bielinskiene@vdu.lt

Abstract

The paper aims at presenting English-Lithuanian corpora for bilingual term extraction (BiTE) in the cybersecurity domain within the framework of the project DVITAS. It is argued that a system of parallel, comparable, and training corpora for BiTE is particularly useful for less resourced languages, as it allows to efficiently use strengths and avoid weaknesses of comparable and parallel resources. A special focus is given to the open nature of the data, which is achieved by publishing the data in CLARIN-LT repository.

1 Introduction

The model of combining several types of corpora has been chosen for the bilingual terminology extraction project DVITAS.[1] The aim of the project is to develop a methodology for automatic extraction of English and Lithuanian terms of a specialised domain from parallel and comparable corpora, as well as to create a publicly available bilingual termbase. Cybersecurity (CS) terminology has been chosen as a specialised domain for the project because of its particular relevance in today's digitalised world in which cyber hygiene skills are indispensable for every Internet user. The compiled termbase is believed to be valuable both for specialists of the domain and the general public, as well as drafters of legal and administrative documents, and translators.

The project aims at employing current deep learning terminology extraction methods. In 2020, the project team (Rokas et al., 2020) completed a pilot study on semi-supervised automatic extraction of Lithuanian CS terms from a Lithuanian monolingual corpus. A small-scale manually annotated dataset (66,706 word corpus with 1,258 annotated cybersecurity terms) was used as training data. The pilot study was performed in several stages: firstly, various baseline LSTM and GRU networks were tested using Adam optimizer and FastText embeddings; secondly, each of the best baseline LSTM and GRU networks were tested with various optimizers; and finally, the best model was compared with a model that has been trained using multilingual BERT embeddings (Rokas et al., 2020) . The latter approach proved to be the most efficient: Bidirectional Long Short-Term Memory model (Bi-LSTM) using multilingual Bidirectional Encoder Representations from Transformers (BERT) embeddings reached F1 score of 78.6%.

We believe that more comprehensive training data obtained from larger manually annotated gold standard corpora will allow to improve the obtained results. In studies by other scholars, who applied neural networks for term extraction as sequence labelling task and used larger annotated datasets, higher F1

[1] https://klc.vdu.lt/dvitas/en

score was achieved: e.g., Kucza et al. used a dataset with 78,567 annotated terms and with Bi-LSTM reached F1 score of 86.73% (Kucza et al., 2018).

The methodology used in the pilot study will be modified and tested on different configurations of neural networks taking into account the methods applied in related research. For instance, studies on sequence labeling tasks with multilingual BERT embeddings show that reduction of the number of languages to three in BERT models may help to achieve higher results compared with the ones achieved with multilingual BERT (Ulčar and Robnik-Šikonja, 2020).

2 Related Research

Bilingual/multilingual term extraction, which is widely used for terminographic purposes, is performed using both parallel and comparable corpora. Term extraction from parallel corpora has been already applied for several decades (Kupiec, 1993). Lately, the importance of comparable data is increasing, as more and more papers have appeared on term extraction from comparable corpora (Vintar, 2010; Delpech et al., 2012; Gornostay et al., 2012; Aker et al., 2013; Chu et al., 2016). Besides, since 2008 the *Workshop on Building and Using Comparable Corpora (BUCC)* has published a number of valuable research papers on the usage of comparable corpora for term extraction.

Researchers indicate several important advantages of using comparable data. Firstly, term extraction from comparable corpora provides valuable terminological data as these data reflect the usage of terminology in original languages which is much more natural than the usage of terminology in translations that are inevitably influenced by source languages. There is a strong possibility of having “inconsistencies in parallel corpora, which are then replicated by translators” (Postolea and Ghivirigă, 2016). Another important advantage of this approach is the possibility to include data sources of a much larger variety as data source search is not limited to translated resources. The third advantage is that comparable corpora are less expensive to build than parallel corpora. The last two are especially important for less-resourced languages which often lack parallel data.

Therefore, some scholars have introduced the idea of combining comparable and parallel corpora to benefit from the advantages provided by both (Bernardini, 2011; Morin and Prochasson, 2011; Biel, 2016; Giampieri, 2018) or yet some researchers concentrate solely on comparable corpora (Steyaert and Rigouts Terryn, 2019; Vintar et al., 2020).

3 Corpora System for Bilingual Terminology Extraction

Five CS corpora have been compiled for this project: a parallel corpus of English texts and their Lithuanian translations (approx. 1.4 million words), a comparable corpus composed of two subcorpora: original English texts and original Lithuanian texts (approx. 4 million words), and three training (gold standard) corpora (approx. 0.1 million words each), which will be manually annotated. The system of corpora and a flowchart of BiTE is presented in Figure 1.

The analysis of the cybersecurity sources revealed that this domain is highly heterogeneous and it encompasses diverse types of information accumulated in various discourses. Ideally, the cybersecurity corpora should be representative of the whole cybersecurity domain and its constituent genres. In order to fully represent the CS domain, we need to consider four discourses as sources of information: legislative & administrative, academic, expert, and popular (the discourses identified by (Wall, 2007)).

Most sources are suitable for compilation of the comparable corpus, which will consist of the original texts in English and Lithuanian. Meanwhile, the sources suitable for the parallel corpus (English original texts and their translations into Lithuanian) are much more sparse.

Legislative and executive sources contain textual information on cybersecurity, such as cybersecurity strategies, laws, government resolutions, minister orders, etc. Official national and EU legally binding and non-binding documents are commonly accessible without any restrictions. The documents of these categories can be acquired for both comparable and parallel corpora (see Table 1 and Table 2).

Another important source of relevant information is scientific research publications on the cybersecurity topic. However, access to academic and scientific publications is often restricted. Most relevant scientific sources are published by major publishing companies and protected by intellectual property

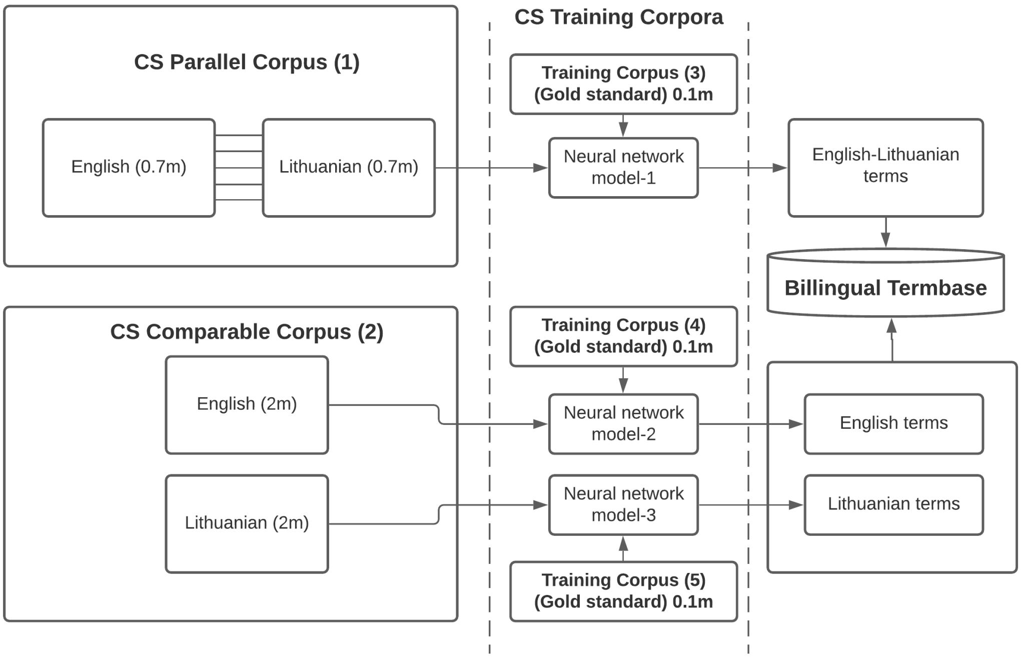

Figure 1: Corpora system for BiTE

rights. As we should ensure proper usage of these texts, the amount of texts suitable for the comparable corpus is rather limited. In fact, it is almost impossible to acquire original and translated academic texts for the parallel corpus.

We had to rely on the inclusion of rather large bulk of media texts into the corpus, due to less-restricted accessibility of these texts. In order to avoid an influx of general terms into the corpus, we tried to include more specialised media sources.

The CS comparable corpus compiled for the project includes texts from the time period of 2010-2020. The categories, subcategories and their approximate proportions within the corpus are presented in Table 1.

Main categories	**Subcategories**	**Proportion**
Legislative and executive documents	CS strategies, laws, government resolutions, minister orders	10%
Official non-binding documents and informational texts	Reports and recommendations of the National Cybersecurity Centres; booklets and posters	15%
Academic texts	Scientific articles, monographs, MA and PhD theses, textbooks	25%
Media texts	Mass media articles, specialised media articles	50%

Table 1: Structure of the comparable corpus (2010-2020).

The parallel corpus includes the EU legal acts and other documents from the time period of 2010-2020. The documents are extracted from the EUR-Lex database and other EU institutional repositories (see Table 2). As mentioned previously, for the parallel corpus it is almost impossible to acquire original and translated academic texts and it is likewise difficult to find translated media articles. Therefore the corpus relies solely on EU documents.

Main categories	Subcategories	Proportion
Legally binding documents (secondary legislation)	Regulations of the European Parliament and of the Council; Directives of the European Parliament and of the Council; Decisions of the European Parliament and of the Council	60%

Official non-binding documents	Communications of the European Commission;
	Reports of the European Commission;
	Recommendations of the European Commission;
	Opinions of the Committees of the EU;
	Briefing papers of the Court of Auditors
		40%

Table 2: Structure of the parallel corpus composed of the EU documents (2010-2020).

Training (gold standard) corpora have been composed of the same text categories as the main corpora. The comparable training corpus encompasses legally biding documents, official non-binding documents and informational texts, academic texts, and media articles. Parallel training corpus is composed of the most important EU legal acts and other documents on cybersecurity issues.

It is important to note that the task of depositing all 5 corpora into CLARIN-LT repository as open access resources is included in deliverables of the project.

4 Concluding Remarks

In the full paper we will present a more lengthy discussion on related research and state-of-the-art approaches to BiTE. Moreover, we will present a detailed discussion on different intellectual property restrictions and how we have dealt with them when compiling parallel and comparable corpora. Finally, the full paper will provide links to the compiled corpora in the CLARIN-LT repository.

Acknowledgements

The research is carried out under the project "Bilingual Automatic Terminology Extraction" funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project is also included as a use case in COST action "European Network for Web-Centred Linguistic Data Science" (CA18209).

References

Ahmet Aker, Monica Lestari Paramita, and Robert J. Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 402–411. The Association for Computer Linguistics.

Silvia Bernardini. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS - A Journal of Professional Communication*, 26.

Łucja Biel. 2016. Mixed corpus design for researching the eurolect: A genre-based comparable-parallel corpus in the pl eurolect project. *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora.*

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2016. Parallel sentence extraction from comparable corpora with neural network features. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Estelle Delpech, Béatrice Daille, Emmanuel Morin, and Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In Martin Kay and Christian Boitet, editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 745–762. Indian Institute of Technology Bombay.

Patrizia Giampieri. 2018. Online parallel and comparable corpora for legal translations. *Altre Modernità*, 20:237–252.

Tatiana Gornostay, Anita Ramm, Ulrich Heid, Emmanuel Morin, Rima Harastani, and Emmanuel Planas. 2012. Terminology extraction from comparable corpora for latvian. In Arvi Tavast, Kadri Muischnek, and Mare Koit, editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 66–73. IOS Press.

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2072–2076. ISCA.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34.

Sorina Postolea and Teodora Ghivirigă. 2016. Using small parallel corpora to develop collocation-centred activities in specialized translation classes. *Linguaculture*, 2016(2):53–72.

Aivaras Rokas, Sigita Rackevičienė, and Andrius Utka. 2020. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In Andrius Utka, Jurgita Vaičenonienė, Jolanta Kovalevskaitė, and Danguolė Kalinauskaitė, editors, *Human language technologies - the Baltic perspective: proceedings of the 9th international conference, Baltic HLT, Kaunas, Vytautas Magnus University, Lithuania, 22-23 September 2020*, pages 39–46. IOS Press.

Kim Steyaert and Ayla Rigouts Terryn. 2019. Multilingual term extraction from comparable corpora: Informativeness of monolingual term extraction features. In Serge Sharoff, Pierre Zweigenbaum, and Reinhard Rapp, editors, *Proceedings of the 12th Workshop on Building and Using Comparable Corpora*, pages 16–25, Varna, Bulgaria, September.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert: less is more in multilingual models. *arXiv e-prints*, June.

Špela Vintar, Larisa Grčić Simeunović, Matej Martinc, Senja Pollak, and Uroš Stepišnik. 2020. Mining semantic relations from comparable corpora through intersections of word embeddings. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 29–34, Marseille, France, May. European Language Resources Association.

Spela Vintar. 2010. Bilingual term recognition revisited. *Terminology*, 16:141–158.

David S. Wall. 2007. *Cybercrime: The Transformation of Crime in the Information Age*. Polity.