Bilingual Terminology Extraction Using Neural Word Embeddings on Comparable Corpora

Darya Filippova, Burcu Can, Gloria Corpas Pastor

Research Group in Computational Linguistics
University of Wolverhampton
University of Malaga

{d.filippova, b.can}@wlv.ac.uk gcorpas@uma.es

Abstract

Term and glossary management are vital steps of preparation of every language specialist, and they play a very important role at the stage of education of translation professionals. The growing trend of efficient time management and constant time constraints we may observe in every job sector increases the necessity of the automatic glossary compilation. Many well-performing bilingual AET systems are based on processing parallel data, however, such parallel corpora are not always available for a specific domain or a language pair. Domain-specific, bilingual access to information and its retrieval based on comparable corpora is a very promising area of research that requires a detailed analysis of both available data sources and the possible extraction techniques. This work focuses on domain-specific automatic terminology extraction from comparable corpora for the English – Russian language pair by utilizing neural word embeddings.

1 Introduction

Terminology studies represent a strictly hierarchical discipline that forms a basement for global scientific knowledge. One of the main challenges of the term extraction task is mostly associated with the "Deficiency of unique norms and rules among linguists and lexicographers to identify a massive amount of modern terminology vocabulary, systematize and place it on international databases because the appearance of new scientific and technological notions occurs faster than it can be defined" (Bidnenko, 2018). Despite the controversy about the notion on "termhood", the main obstacle in the term extraction process consists in the restricted access to cross-lingual parallel content being a major bottleneck for compiling a high-quality specialised dictionary with minimum effort.

Many well-performing bilingual AET systems are based on processing parallel data, however, such parallel corpora are not always available for a specific domain or a language pair (Uyeama, 2006). Even though the number of parallel corpora available online is gradually increasing, this number is still not sufficient to include a wide range of all possible domains, especially if the domain is relatively new. Comparable corpora, on the contrary, can be an excellent alternative with a lot of potential and greater possibilities for domain- and language adaptability (Pekar et al., 2008). Other advantages of comparable corpora include its relatively cheap construction process and its usefulness in building a domain-specific corpus even for low-resource languages and underrepresented topics. It is important to note that a specialised comparable corpus is traditionally of a smaller size (when compared to a general-language corpus), this is usually related to the scarcity of resources in languages other than English (Hazem and Morin, 2016). Such a corpus may also pose issues associated with the unfavorable duplication of web-crawled texts, noisy data, and a general lack of control over the corpus compilation process. Nevertheless, the use of comparable corpora remains a very promising area of research in both monolingual and bilingual term extraction.

This research is hypothesized on the idea that language specialists are usually provided with the minimum information about an event and little or no additional information about the specific domain they must work with (Gaber et al., 2021). The general pipeline of the terminology extraction process can be divided into three main stages: 1. Building a corpus; 2. Application of term extraction methods; 3. Either automatic or

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 58–64,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_009

manual validation of extracted term candidates. Considering all the advantages of using a bilingual comparable corpus, we would like to exploit its properties in the terminology extraction task and make it domain-independent i.e., adaptable to any other domain where the compilation of corpora is possible; However, this work will be mainly focused on Covid – 19 domain whereas our rationale will be discussed later in the next sections

2 Related Work

2.1 Statistical Term Extraction Methods

The variety of strategies used for the automatic term extraction was developed with one common objective to obtain the most consistent set of terms being the best representation of a domain in question. The commonly agreed approaches used in the task of term extraction can be clustered into three main groups, depending on the methods of scoring the candidate terms (Astrakhantsev, 2018). Thus, these three main categories are also classified as statistical, linguistic, and hybrid, accordingly. Statistical approach was aimed to determine the potential "termhood" [1] of a candidate word by defining the optimum measures of this "termhood" (Pazienza et al., 2006). The history of frequency-based approaches lasts for decades, thus Total TF-IDF algorithm is based on measuring word informativeness, i.e., the importance of a particular word/potential term in a given document (Evans and Lefferts, 1995). C-value is another traditional yet more comprehensive statistical measure, that is commonly used with multiword expressions, this measure does not only consider the frequency of an expression but its length and the frequency of its constituents in a corpus, it is proved to be very efficient when applied to highly technical domains (Frantzi et al., 2000). Mutual Information (MI) is widely used as one of the statistical measures of term frequency where the independent probability of terms is compared to the probability to see these terms together. MI is very useful for identifying high-frequency terms; however, it overestimates multiword terms consisting of rarely seen words (Church et al., 2003). The idea of using frequency information as the feature for identifying terms has not only resulted in specialized scripts but also inspired the development of frequency-based automatic term extractors. One of such tools is an open-source TBXTools[2] developed by A. Oliver and M. V`azquez (2015) with the aim of automating statistical term extraction process and making it realizable for general users with no advanced programming skills. With reference to other language-dependent methods, an attempt to compile a glossary of scientific terms in Russian was made by Bolshakova et al., (2019). Considering the complicated structure of Russian grammar, a set of lexical and syntactic patterns was formed to detect terms based on syntactic and morphological features of the Russian language. Such a collection of rules is very efficient when extracting terms in a particular domain, however a very specific fine-tuning is needed depending on every domain, its size, and language.

2.2 Linguistic and Hybrid Term Extraction

Linguistic methods of term extraction try to identify terms based on their syntactic features. Scholars supporting this approach state that syntactic information is sufficient to determine the termhood of a word (Bourigault, 1992). In their comprehensive review Pazienza et al., (2006) confirm that these two stages can be sufficient for loose term extraction if the initial POS grouping is performed correctly and is based on the deep analysis of the most common syntactic structures of a language in question. Let us consider such a common phrase as "*Health protection*", its equivalent in Russian is "*Охрана здоровья*", however, the chances that we will be able to obtain the same translation equivalent using the frequency-based or linguistic techniques are quite low, as this term could potentially be used in different contexts, thus the declension of the first word may change to genitive case ("*Охраны здоровья*") or dative case ("*Охране здоровья*") and others depending on the context. It is important to note that within a linguistic framework of AET other refinement techniques can be applied for further term filtering, such as the creation of customized lists of unwanted words, general-purpose words, or

[1] Termhood is defined as a potential ability of a word to be a term

[2] https://github.com/aoliverg/TBXTools

functional words followed by the manual validation of subject-matter experts (Pazienza et al., 2006).

The best practice in the traditional approach to AET is to use a combination of both statistical and linguistic techniques to achieve better performance. Such hybrid methods are commonly used in the automated term extraction systems. For example, TermoStat [3] tool was developed with the aim of extracting corpus-specific terminology and performed quite well when extracting simple terms achieving 74% of precision (Drouin, 2003). Another hybrid software used for both term extraction and corpus compilation is SketchEngine [4] . This software supports a range of languages and can even be used for domain-specific bilingual term extraction if a parallel corpus is available (Kilgarriff et al., 2014). Some other tools combine nearly all available traditional approaches (ATR4S software that comprises 15 AET methods, including, but not limited to Average Term Frequency, Residual IDF, Relevance and Weirdness scores, etc.) and deliver good results (Astrakhantsev, 2018). Even though the aforementioned techniques performed quite well in both monolingual and bilingual environment, the growing interest to state-of-the-art machine learning methods led to numerous experiments carried out in a range of NLP tasks including term extraction.

2.3 Experiments With the Word Embedding Techniques

The main impulse in improving the consistency and accuracy of domain-specific terminology sets was to increase the number of features each term can be represented with rather than treating every word as an atomic unit (Amjadian et al., 2016). With the introduction of word embeddings, it became possible to obtain deeper contextual meanings of words being a set of features. Tomas Mikolov (2013) was the first to propose "A method to learn a linear transformation from the source language to the target language" to improve the task of lexicon extraction from bilingual corpora. Thus, Hazem and Morin (2017) attempted to improve bilingual terminology extraction from specialised corpora by exploiting finer granularity of the word embeddings generated by Word2Vec. In their work, it was argued that the traditional approaches are also not suitable for domain-specific term extraction as the word co-occurrence information for a small size comparable corpus is always not reliable as well as pre-trained neural network models are unable to learn the features if the vocabulary is of modest size. This limitation was then addressed by enriching the distributed word representations with the general domain data. The concatenation of distributed word representations trained on a domain-specific corpus and a general corpus helped to improve the quality of extracted terms comparing to previous works where the algorithms were only trained on specialised corpora.

It is important to mention that the introduction of other word embedding architectures influenced further experiments with vector concatenation. So, Amjadian (2016) used the abovementioned CBOW and Skip-gram architectures to produce so-called "global word vectors" trained on a general language corpus and concatenated them with "local context vectors" generated by means of another word embedding architecture called Glove (Pennington et al., 2014). Besides the promising results of the previously described works, there are some other limitations, that were not taken into account, such as the ability to produce bias results due to the unbalanced dimensionality of domain-specific and general word vectors. This limitation was then addressed in a later work of Liu et al., (2018) where they proposed to restrict the size of domain-specific word vectors to 100 comparing to the size of general domain word vectors being set as 300. This fine-tuning allowed to not only preserve the features of both corpora but to also use available pre-trained models with the minimum changes required.

Even though multiple experiments with the various types of word embeddings for inflected and agglutinative languages (Üstün et al., 2018; Romanov and Khusainova, 2019) have shown that morphological subword embeddings perform very well with the Slavic languages, these types of distributed word representations are still limited due to its static nature i.e., inability to change depending on the context. However, more

[3] http://termostat.ling.umontreal.ca/

[4] https://www.sketchengine.eu/

information extraction opportunities occurred with the introduction of the most modern type of deep contextualized word embedding, such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019). Among other possible applications, the BERT model was also considered in the context of binary classification for associating term candidates with their context (Hazem et al., 2020). Given the wide range of neural network-based term extraction experiments we are convinced that state-of-the-art methods can help to improve the AET task for the English – Russian language pair by either being paired with the traditional methods or by applying modern vector concatenation techniques.

3 Methodology on Automatic Term Extraction and Bilingual Mapping of the Terms

3.1 Comparable Corpora as the Main Source of Data

A good example of data scarcity case can be demonstrated by the availability of parallel datasets for En-Ru language pair on such topics as Coronaviruses and Vaccination. The rapid growth of interest to this topic is disproportional comparing to the amount of available bilingual data. As of July 2021, the only open-source parallel corpus we could find on the web was Corona Crisis Corpora published by TAUS[5] with the aim of ensuring better language coverage. The corpus consists of 192,614 aligned segments and could be very useful for small-scale NLP experiments, however, this data is still not enough to cover all Covid-related domains.

Numerous works in the sphere of term extraction from specialized corpora did not only differ in methods and techniques but also covered many domains and languages. Several factors were considered while choosing a domain for our datasets, which is:

a. Asymmetrical availability of resources – by focusing our experiment mostly on the topic of Covid-19, we believe that we will be able to contribute to the simplification of term extraction process for such domains, where the information in English is prevalent due to their novelty. The topic of the coronavirus remains one of the most popular areas of modern research, however, most up-to-date scientific papers on Covid – 19 are published in English. Consequently, the number of relatively new terms introduced in this language is much greater than it is in any other high-resource languages. This disproportion of terminological equivalents makes the task of glossary compilation even more challenging.

b. Usefulness and timeliness of the topic – despite its novelty, Covid – 19 led to numerous international events where the assistance of language specialists is crucial. We are convinced that this domain will not only remain in high demand for the next several years but will also give rise to the development of other cross-disciplinary domains where the availability of high-quality bilingual terminology sets will be essential.

The comparable corpus of the En-Ru language pair will consist of two monolingual corpora, where the English part of it will be represented by the CORD-19 dataset [6] . This exhaustive, open-source dataset was compiled by the leading research groups of the world as a response to the COVID-19 pandemic and it consists of more than 500 000 scientific articles on SARS-COV-19. The compilation of a corpus in Russian will be performed by the automatic web crawling with the subsequent data cleaning and processing. Several websites will be considered as the main source of data in Russian to ensure that the dataset is represented by scientific papers containing plenty of terms. We would like to follow the methodology of A. Rungsawang (2004) for building the web crawler and to ensure that the information from the web pages is collected in an incremental way to avoid data duplication during the subsequent crawls. Thus, the crawler algorithm will be able to learn from search experience and produce filtered results in case of several crawling attempts. Topic relevancy score will be used together with the boilerplate removal to form the textual data. Besides the possibility to have noisy data that requires special treatment and extra attention, another debatable peculiarity of a comparable corpus remains its "comparability" or similarity of genres of the chosen subcorpora,

[5] https://md.taus.net/corona

[6] The dataset is available at: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

therefore the initial selection of data sources will be of paramount importance.

3.2 Word Embeddings as Input to a Neural Network Classifier

Word vectorization is performed by means of pre-trained word embedding models to minimize the effort of parameter engineering and benefit from state-of-the-art algorithms. The final optimal approach will solely depend on the results of the experiments carried out with the concatenation of distributed word representations and changing the dimensionality of word embeddings as well as associating specialized corpora with the external data. There is a range of approaches to using bilingual neural network classifiers to distinguish between the terms and non-terms. Some of such deep learning algorithms are trained on a small set of labelled data and then tested on predicting the class on the big amount of domain-specific unlabelled data (Wang et al., 2016). Hand engineering is a very time-consuming and labor-intensive task that often requires extra supervision and the assistance of subject-matter experts. Following the trend of minimum hand–engineering, another term extraction approach relies on the benefits of using deep contextualized word embeddings as input to Convolutional Neural Networks with minimum supervision (Khosla et al., 2019). We would like to adopt and combine the previous methods proposed by Hazem and Morin (2017), Amjadian et al., (2016), and Liu et al., (2018) and to exploit the idea of the independent training of local and global vectors with minimum fine-tuning of more modern customizable models that are capable of learning the context of words.

3.3 Bilingual Mapping and Results Evaluation

Considering the monolingual training conditions, one of the final stages of our research will be to find the translation equivalents for the set of candidate terms extracted by the classifier. The task of bilingual mapping, as well as various transformation options, were thoroughly studied by Artetxe et al., (2016) where they proposed a framework for learning the optimal vector transformation that outperformed all the previous techniques and helped to solve the task of zero-shot translation[7]. We would like to follow the upgraded version of the suggested multi-step framework proposed by Artetxe et al., (2018) consisting of the orthogonal mapping of the word vectors and reducing their dimensionality. This technique proved to be very efficient when paired with the state-of-the-art word embeddings trained on two corpora of distant language pairs.

As it was mentioned before, considering the uncertainties about the classification and the definition of terms, the task of data validation in this work is quite challenging. It was decided to avoid human evaluation techniques because a.) we are convinced that this task is very subjective for manual evaluation as the entire process of glossary building often depends on personal preferences and the individual level of expertise b.) this process is very labor-intensive and requires the assistance of domain experts that are not always available. Alternately, standard automatic evaluation metrics such as precision, recall, and F score will be used. To ensure proper evaluation of the terminology extraction task the terminology reference list for the En-Ru language pair is required. It is important to mention that the domain of Covid-19 is still a field of active research and new terminology is coined every day, however, we could not find any gold standard datasets that are currently available on the web. Thus, the reference bilingual terminology list will be compiled manually by consolidating several glossaries provided by such organizations as WIPO Pearl [8] and the official terminology database of the European Union [9]. The list of possible reference sources is not exhaustive and will be updated in the course of searching for the most up-to-date open-source termbases published by trustworthy providers.

4 Research Questions

The main motivation of this research relates to the restricted access to cross-lingual content being a major obstacle for compiling a high-quality specialised dictionary with minimum effort.

[7] Zero-shot translation enables the neural network to pick the nearest possible translation equivalent for the words that were not present in the training data set.

[8] COVID 19 multilingual glossary is available at: https://www.wipo.int/pressroom/en/articles/2020/article_0021.html

[9] https://data.europa.eu/data/datasets/covid-19-multilingual-terminology-on-iate?locale=en

Thus, our research questions can be formulated as follows:

- What are the limitations of generating a bilingual dictionary for a specific domain when no additional resources are provided?
- Is it possible to compile bilingual glossaries from bilingual comparable corpora?
- If so, what is the most efficient procedure to perform bilingual term extraction?

It is important to note that we would like to address the realia of working conditions that usually imply a significant amount of information that has to be processed by an interpreter in a short period of time to manually compile a glossary; hence our main goal will be to automate glossary creation process to the maximum possible extent and, at the same time, preserve the accuracy and consistency of specialised terms.

References

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

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, pages 2289–2294. https://doi.org/10.18653/v1/D16-1250

Artetxe, M., Labaka, G., & Agirre, E. (2018). Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012-5019.

Astrakhantsev, N. (2018). ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala. *Language Resources and Evaluation, 5 2*, pages 853-872.

Bidnenko, N. (2018). *Practical and theoretical issues of modern terminology* [Ph.D. Thesis, Alfred Nobel University, Dnipro], pages 3-5.

Bolshakova, E., Efremova, N., & Ivanov, K. (2019). Terminological Information Extraction from Russian Scientific Texts: Methods and Applications. pages 7-11. https://doi.org/10.29007/k93q

Bourigault, A. (1992). Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. *COLING 1992 Volume 3: The 14th International Conference on Computational Linguistics*, https://www.aclweb.org/anthology/C92-3150

Church, K., Gale, W., Hanks, P., & Hindle, D. (2003). Using Statistics in Lexical Analysis. *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages. 5-9.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pages 4171 – 4186.https://doi.org/10.18653/v1/N19-1423

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), pp.99–115. https://doi.org/10.1075/term.9.1.06dro

Evans A. and R. G. Lefferts. (1995). CLARIT-TREC experiments. In Information Processing and Management: An International Journal, volume 31, pages 385–395.

Frantzi K., Ananiadou S., and Mima H. 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. International Journal on Digital Libraries, pages 115–130.

Gaber, M., Corpas Pastor, G., & Omer, A. (2021). Speech-to-Text Technology as a Documentation Tool for Interpreters: a new approach to compiling an adhoc corpus and extracting terminology from video-recorded speeches. *TRANS. Revista de Traductologia*, pages 263-281. https://doi.org/10.24310/TRANS2020.v0i24.7876

Hazem, A., & Morin, E. (2016). Efficient Data Selection for Bilingual Terminology Extraction from Comparable Corpora. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pages 3401–3411.

Hazem, A., & Morin, E. (2017). Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Taipei, Taiwan, pages 485-493.

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

https://www.aclweb.org/anthology/2020.computerm-1.13

Khosla, K., Jones, R., & Bowman, N. (2019). Featureless Deep Learning Methods for Automated Key-Term Extraction. Department of Mathematics. Stanford University. Stanford, pages 2-8.

Kilgarriff, A., Baisa, V., Busta, J., Jakubícek, M., Kovár, V., Michelfeit, J., Rychlý, P., & Suchomel, V.
(2014). The Sketch Engine: ten years on. *Lexicography, pages* 7-36.

Liu, J., Morin, E., & Saldarriaga, P. (2018). Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pages 2855–2866.

Mikolov, T., Quoc, V. L. e., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. v.1, pages 2-4. https://doi.org/https://arxiv.org/abs/1309.4168

Oliver, A., & Vàzquez, M. (2015). TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pages 473–479. https://www.aclweb.org/anthology/R15-1062

Pazienza, M., Pennacchiotti, M., & Zanzotto, F. B. (2006). Terminology extraction: an analysis of linguistic and statistical approaches. *Knowledge Mining pp.255-279*, pages 255-279. https://doi.org/10.1007/3-540-32394-5_20

Pekar, V., Mitkov, R., Blagoev, D., Mulloni, A. (2008). Finding Translations for Low-Frequency Words in Comparable Corpora. *Machine Translation*, 20 (4), pages 247-266

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pages 2227–2237. https://doi.org/10.18653/v1/N18-1202

Romanov, V., & Khusainova, A. (2019). Evaluation of Morphological Embeddings for English and Russian Languages. *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, Minneapolis, USA, pages 77-81.

Rungsawang, A., & Angkawattanawit, N. (2004). Learnable topic-specific web crawler. *Department of Computer Engineering, Massive Information & Knowledge Engineering*, Bangkok, Thailand, pages 5-12.

Üstün, A., Kurfalı, M., & Can, B. (2018). Characters or Morphemes: How to Represent Words?. *Proceedings of the 3rd Workshop on Representation Learning for NLP*, Melbourne, Australia, pages 144-153.

Uyeama, M. (2006). Building general- and special-purpose corpora by Web crawling. *The 13th NIJL International Symposium, Language Corpora: Their Compilation and Application.*, Tokyo, Japan, pages 2-6.

Wang, R., Liu, W., & McDonald , C. (2016). Featureless Domain-Specific Term Extraction with Minimal Labelled Data. *Proceedings of the Australasian Language Technology Association Workshop 2016*, Melbourne, Australia, pages 103–112. https://www.aclweb.org/anthology/U16-1011.pdf