Michaela Denisová, Gilles-Maurice de Schryver, and Pavel Rychlý

THE AUTOMATIC DETERMINATION OF TRANSLATION EQUIVALENTS IN LEXICOGRAPHY: WHAT WORKS AND WHAT DOESN'T?

Abstract Cross-lingual embedding models act as facilitator of lexical knowledge transfer and offer many advantages, notably their applicability to low-resource and non-standard language pairs, making them a valuable tool for retrieving translation equivalents in lexicography. Despite their potential, these models have primarily been developed with a focus on Natural Language Processing (NLP), leading to significant issues, including flawed training and evaluation data, as well as inadequate evaluation metrics and procedures. In this paper, we introduce cross-lingual embedding models for lexicography, addressing the challenges and limitations inherent in the current NLP-focused research. We demonstrate the problematic aspects across three baseline cross-lingual embedding models and three language pairs and outline possible solutions. We show the importance of high-quality data, advocating that its role is vital compared to algorithmic optimisation in enhancing the effectiveness of these models.

Keywords translation equivalent determination; cross-lingual embedding models; evaluation

1. Introduction

In computational lexicography, translation equivalent retrieval is a non-trivial task with many challenges, including language ambiguity, variety, and disparity. Pinpointing translation equivalents relies heavily on human intuition and is time-consuming. An alternative approach is to use fully-automated Natural Language Processing (NLP) methods and algorithms, speeding up the process by offering a set of translation-equivalent candidates for a headword.

Cross-lingual embedding models represent a category of methods suitable for this task. These models are appealing and beneficial for lexicography for the following reasons. Firstly, as opposed to their parallel-data-based counterparts, they utilise *comparable data*, which are thus also available for under-resourced and non-standard language combinations. Secondly, the text variety of comparable data is usually much wider than in parallel corpora, offering more natural equivalents and vocabulary.

Despite these advantages, the research that has been done on these models is typically NLP-focused, improving the optimisation of the algorithms and technical aspects of these models while neglecting the lexicographic point of view (Ruder et al., 2019). The problems that arise from this, range from incorrect terminology usage to erroneous training and evaluation data and inappropriate evaluation metrics and methods, impeding our ability to interpret the results correctly and monitor the progress accurately.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

In this paper, we introduce cross-lingual embedding models for translation equivalent retrieval in lexicography. We provide new insights about the training data, evaluation data, methods and metrics. We show how the selection of training data influences the results, analyse the widely used evaluation datasets and assess evaluation methods and metrics on three example language pairs: English-Slovak, English-Estonian, and English-Korean. We point out the problems and challenges and outline possible solutions. We show why the currently-used training data and evaluation data are faulty, why the evaluation method and metrics are inadequate, and how each of these can be improved.

Our motivation is to provide valuable material for lexicographers about cross-lingual embedding models for translation-equivalent retrieval that could serve as a stepping stone in their bilingual dictionary projects. Moreover, we emphasise the importance of high-quality data, which is often more crucial for achieving desired outcomes than the optimisation of the underlying algorithm.

The article is structured as follows. Section 2 outlines the background of cross-lingual embedding models. Section 3 compares two sources of the training data and demonstrates the results across three baseline models. Section 4 analyses evaluation datasets for three language pairs and shows how to allow for inaccuracies in them during evaluation. Section 5 discusses the most common evaluation metric and suggests an alternative. Section 6 offers concluding remarks.

2. Cross-Lingual Embedding Models

Word embeddings, or vector representations of words, are used to mathematically express a word's meaning based on the context words accompanying the word (Smith, 2019). In recent years, cross-lingual embedding models have appealed to many researchers mainly due to their ability to connect meanings across languages. These models aim to align two (or more) sets of separately trained monolingual word embeddings into a shared space where similar words obtain similar vectors (Ruder et al., 2019).

Except for monolingual word embeddings, they could involve a certain level of supervision in training. Based on the size of the training datasets in a word-to-word format, the models can be categorised into supervised (usually up to 5,000 headwords in the training dataset), semi-supervised (a small training dataset or mode that relies on identical strings and numerals occurring in monolingual word embedding vocabulary), and unsupervised (no training dataset) (Ruder et al., 2019).

In NLP, translation equivalent retrieval is referred to as the bilingual lexicon induction task (BLI) or the bilingual dictionary induction task (BDI). From a lexicographic point of view, this terminology is incorrect since "lexicon" or "dictionary" have much broader meanings and are not limited to translation equivalents only.

In the BLI task, the main objective is to find the most suitable translation equivalent(s) for a headword, typically by computing cosine similarity between the vectors of the

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

headword and the translation-equivalent candidates. Afterwards, the retrieved list of translation-equivalent candidates is compared to the evaluation dataset, a word-to-word list (Ruder et al., 2019).

Additionally, precision at *k* (P@*k*) is the most common reported metric, where *k* represents the number of translation-equivalent candidates retrieved for a headword. The precision computes the ratio of the correctly retrieved translation equivalents to all retrieved translation-equivalent candidates. Other crucial metrics, such as recall and F1 score, are less frequently employed. Recall measures the ratio of the correctly retrieved translation equivalents to the word pairs occurring in the evaluation dataset, while the F1 score captures both metrics, precision and recall, showing the balance between them.

In this paper, we selected the three most cited baseline cross-lingual embedding models for our experiments: MUSE (Conneau et al., 2018), VecMap (VM) (Artetxe et al., 2018a; b), and RCLS (Joulin et al., 2018). All three models are trained in a supervised mode (MUSE-S, VM-S, RCLS), whereas only the MUSE and VM models are additionally trained in an unsupervised mode (*model*-U) and a mode that relies on identical strings and numerals (*model*-I).[1]

3. Training Data

Most cross-lingual embedding models are rooted in monolingual word embeddings, or they utilise them directly in the training process (Ruder et al., 2019). The prevailing monolingual word embeddings exploited in the majority of papers (Ren et al., 2020; Mohiuddin et al., 2020; Marchisio et al., 2022; Sanigrahi & Read, 2022; etc.) are pre-trained FastText embeddings (Grave et al., 2018).

The FastText embeddings were trained with dimension 300 on the Wikipedia corpus. They are noisy and have limited vocabulary in size and quality. The vocabulary ranges from 300,000 to 800,000 word forms depending on the language and is often polluted with English words or words from other languages. After examining pre-trained Estonian, Slovak, and Korean FastText embeddings, we discovered many English words, for instance, *align*, *score*, *right* (Slovak embeddings), *Norwegian*, *consciousness*, *history* (Estonian embeddings), *and, border*, *left* (Korean embeddings) considered as high-frequency words in these languages. The reason for this might not stem only from the data itself but also from the data processing method.

The research in cross-lingual word embeddings has shown that the resulting embeddings' quality heavily depends on the monolingual word embeddings utilised during training (Vulić et al., 2020). Therefore, in this paper, we examine how the different sets of high-quality monolingual word embeddings impact the performance of the models. We compare the common FastText embeddings to the less noisy Sketch Engine embeddings (Herman, 2021). The Sketch Engine embeddings were trained with dimension 100 using the same algorithm as FastText but on a different data

[1] RCLS model is available in the supervised mode only.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

source – web corpora, with the vocabulary ranging from two to five million word forms. Table 1 displays the results.

Table 1: Reported precision@1 (P@1) for baseline models trained using FastText (FT) and Sketch Engine (SE) monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

P@1	en-sk		en-et		en-ko	
	FT	SE	FT	SE	FT	SE
MUSE-S	30.53	**37.00**	23.87	**29.6**	**25.73**	17.47
MUSE-I	29.87	**33.40**	23.13	**24.87**	**17.61**	9.42
MUSE-U	19.07	**34.53**	0.0	**21.73**	0.0	0.0
VM-S	35.93	**41.20**	**33.80**	32.33	**35.29**	22.05
VM-I	36.60	**36.67**	**27.67**	23.80	**22.12**	10.92
VM-U	33.13	**37.00**	**25.53**	24.93	**17.34**	10.03
RCLS	38.67	**43.40**	33.87	**38.13**	**35.63**	28.60

When looking at Table 1, we can see that using high-quality monolingual word embeddings, Sketch Engine enhances the performance of all models trained on English-Slovak language pairs and a (slight) majority of the models trained on English-Estonian, within a margin of approximately 0.07% to 15%. The exception is the English-Korean language pair, where models using FastText monolingual word embeddings outperformed all those using Sketch Engine, within a margin of up to 13%.

This demonstrates that not only training data plays a pivotal role but also the quality of the evaluation datasets. In this case, we utilised the MUSE evaluation datasets which, as we show in Section 4, contain many errors, including English-to-English word pairs, predominantly occurring in the English-Korean dataset. This artificially increased the performance of the models trained with FastText monolingual embeddings, which are also heavily polluted with English words. In the following Section, we analyse the evaluation datasets and propose a solution to account for the errors during the evaluation.

4. Evaluation Datasets

The evaluation datasets are compiled based on the vocabulary of the monolingual word embeddings exploited during the training, usually employing an automatic method, such as the popular and widely used MUSE evaluation datasets (Conneau et al., 2018) published along with the eponymous baseline model. These datasets have often been criticised for disproportional part-of-speech distribution, with most word pairs consisting of proper names, which are unsuitable for assessing the quality of the retrieved translation equivalents (Kementchedhjieva et al., 2019). However, as our analysis reveals, these datasets contain many other errors.

In this paper, we analyse the errors in three evaluation datasets: English-Slovak, English-Estonian, and English-Korean[2] and categorise them. Table 2 outlines the

[2] The analysis of the English-Korean evaluation dataset focuses on the English headwords only.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

most common error types occurring in the evaluation datasets. Figures 1-3 illustrate the number of word pairs in each category for each evaluation dataset.

Table 2: Error categories, their description, assigned weight, and an example of a word pair from the respective category

Category	Description	Weight	Example
A	Correct	1	'*accidentally*' : '*náhodne*' (sk)
B	Inflected word form	0.8	'*workshops*' : '*seminaaride*' (*seminaarid*) (et)
C	Proper name	0.3	'*Ahmeď*', '*Bruno*', '*Julia*'
D	Part-of-speech mismatch	0.2	'*darkness*' : '*temné*' (*temnota*) (sk)
E	English to English	0.2	'*cage*' : '*cage*'
F	Abbreviation	0.2	'*cnn*', '*fbi*', '*ibm*'
G	Synonym	0.7	'*customer*' : '*odberateľ*' (*zákazník*) (sk)
H	Incomplete word	0.3	'*cooperative*' : '*ühistu*' (*cooperative society*) (et)
I	Interjection	0.6	'*boom*', '*bang*'
J	Missing diacritics	0.4	'*everybody*' : '*kazdy*' (*každý*) (sk)
K	Incorrect	0.1	'*lucky*' : '*veab*' (õnnelik) (et)

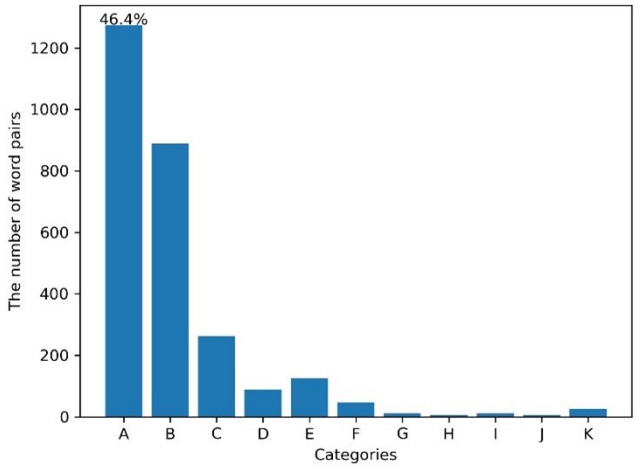

Fig. 1: Number of word pairs in each category in the English-Slovak evaluation dataset

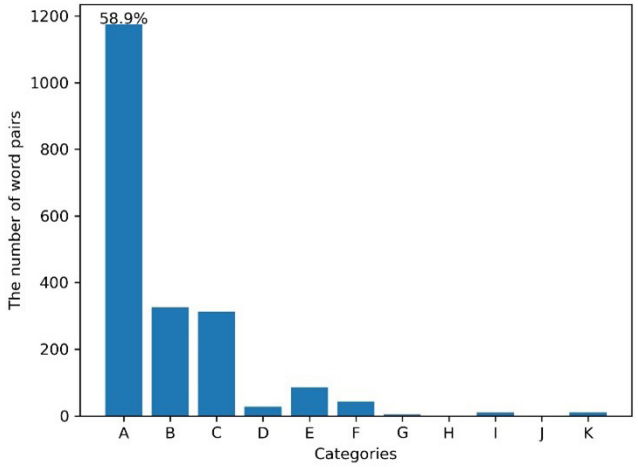

Fig. 2: Number of word pairs in each category in the English-Estonian evaluation dataset

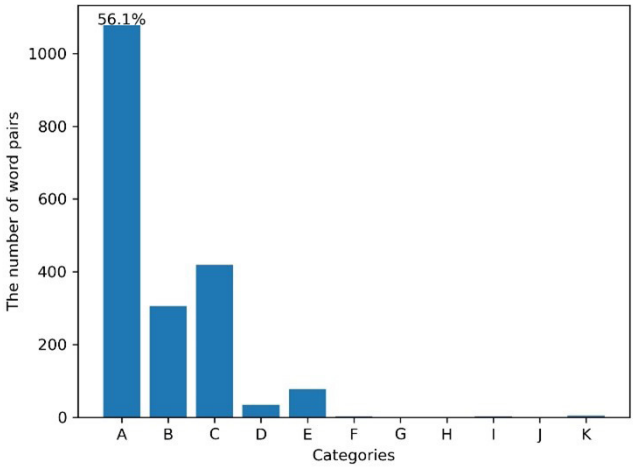

Fig. 3: Number of word pairs in each category in the English-Korean evaluation dataset

According to Table 2, there are 11 types of errors with different levels of severity, such as part-of-speech mismatch (e.g., English-Slovak dataset: *darkness* (noun) – *temné* (adjective, *dark*)), inflected word forms (e.g., English-Estonian dataset: *skill* (sg.) – *oskused* (pl., *skills*) and same-word translations (e.g., English-Slovak dataset: *galaxy* – *galaxy*, English-Korean dataset: *bet* – *bet*), etc. Moreover, given Figures 1–3, we can observe that nearly half of the word pairs in each evaluation dataset contain some error. Such erroneous datasets can distort the actual outcomes.

However, we can allow for these inaccuracies during the evaluation by weighting the precision based on the reliability of the current word pair from the evaluation dataset. In the next step, we assign a weight to each category based on our judgement of the severity of the error, ranging from 1 (the most reliable) to 0.1 (the least reliable). Another reason for weighting is to increase the performance when the model is able to find a word pair with a higher weight and not penalise the model when it does not include a word pair with a low weight. Weights for each category are listed in Table 2. Table 3 shows how the results change when weights are applied.

Table 3: Reported non-weighted (NW) and weighted (W) precision@1 (P@1) for baseline models trained using FastText monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

P@1	en-sk		en-et		en-ko	
	NW	W	NW	W	NW	W
MUSE-S	**30.53**	26.09	**23.87**	21.61	**25.73**	18.81
MUSE-I	**29.87**	24.28	**23.13**	17.69	**17.61**	12.82
MUSE-U	**19.07**	16.34	**0.0**	0.0	**0.0**	0.0
VM-S	**35.93**	29.52	**33.80**	28.73	**35.29**	27.11
VM-I	**36.60**	29.25	**27.67**	22.61	**22.12**	15.90
VM-U	**33.13**	28.11	**25.53**	21.26	**17.34**	13.69
RCLS	**38.67**	33.28	**33.87**	30.13	**35.63**	27.86

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Based on Table 3, when weights are applied, the precision decreases by around 2% to 8%, while the gaps are closer to the upper limit for the English-Korean language combination. This signifies that the English-Korean evaluation dataset contained more severe error types, and the models were better at finding such word pairs.

On top of that, adjusting the weights altered the rankings of the models' performance. For example, the VM-I model, evaluated on the English-Slovak language pair, initially outperformed the VM-S model when no weights were applied. However, upon introducing weights, the VM-S model demonstrated superior performance.

5. Evaluation Metric

Another problematic aspect of the evaluation is the employed metric. The most reported metric is precision at *k* (P@*k*), where *k* represents the number of translation-equivalent candidates retrieved for a headword. In many papers, the *k* is set to 1, 3 or 5, i.e., fixed. However, the number of translation equivalents for each headword differs from word to word and dataset, including the MUSE evaluation datasets. Table 4 shows the statistics of the number of translation equivalents in the English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko) evaluation datasets.

Table 4: Number of translation equivalents (TEs) in three MUSE evaluation datasets

TEs	en-sk	en-et	en-ko
1	760	1111	1085
2	395	303	310
3	208	71	63
4	88	14	7
5+	40	1	0

Table 4 confirms that having a fixed *k* is not an appropriate setting for an evaluation. Additionally, reporting purely on precision is not informative about the actual model's performance. Denisová & Rychlý (2023) showed that the higher the precision the model obtains, the lower the recall. This is also caused by fixed *k*, which often imbalances precision and recall. Instead of preferring one or the other, we can implement a simple classification neural network[3] which classifies the word pairs found by the cross-lingual embedding models as correct or incorrect while allowing for a dynamic *k*. The aim is to identify as many relevant translation equivalents as possible for each headword and report precision, recall, and F1 scores not constrained by a predefined set of *k* translation-equivalent candidates while balancing precision and recall and maintaining high performance. Tables 5, 6, and 7 outline the results of precision, recall, and F1 scores evaluated using fixed and dynamic *k*.

[3] The technical details of the classification neural network and an extensive evaluation across a wide range of language pairs will be published in a separate paper.

Table 5: Reported precision (P), recall (R), and F1 score using fixed k (k = 1) and dynamic k (NN) for the baseline model VM-S trained using FastText monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

VM-S	P@1	P@NN	R@1	R@NN	F1@1	F1@NN
en-sk	**35.93**	34.83	19.68	**44.30**	25.43	**39.00**
en-et	33.80	**36.13**	25.46	**42.16**	29.05	**38.91**
en-ko	35.29	**39.06**	26.90	**45.45**	30.53	**42.02**

Table 6: Reported F1 scores using fixed k (k = 1) and dynamic k (NN) for baseline models trained using FastText monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

	en-sk		en-et		en-ko	
	F1@1	F1@NN	F1@1	F1@NN	F1@1	F1@NN
MUSE-S	21.61	**32.14**	**20.51**	20.18	22.26	**30.86**
MUSE-I	21.14	**29.59**	19.88	**33.03**	**15.23**	15.22
MUSE-U	13.49	**34.44**	0.0	0.0	0.0	0.0
VM-S	25.43	**39.00**	29.05	**38.91**	30.53	**42.02**
VM-I	25.90	**33.00**	23.78	**30.77**	**19.13**	18.54
VM-U	23.45	**41.72**	21.94	**32.77**	15.00	**22.05**
RCLS	27.36	**36.02**	29.10	**33.03**	30.82	**41.07**

Table 7: Reported F1 scores using fixed k (k = 1) and dynamic k (NN) for baseline models trained using Sketch Engine monolingual word embeddings across three language pairs: English-Slovak (en-sk), English-Estonian (en-et), and English-Korean (en-ko)

	en-sk		en-et		en-ko	
	F1@1	F1@NN	F1@1	F1@NN	F1@1	F1@NN
MUSE-S	26.19	**45.11**	25.44	**40.00**	15.12	**23.62**
MUSE-I	23.64	**43.92**	21.37	**29.17**	8.15	**20.83**
MUSE-U	24.44	**41.44**	18.68	**32.89**	0.0	0.0
VM-S	29.16	**47.59**	27.79	**39.63**	19.07	**33.09**
VM-I	25.95	**42.76**	20.45	**27.81**	9.45	**18.93**
VM-U	26.19	**46.98**	21.43	**36.97**	8.68	**13.79**
RCLS	30.71	**51.74**	32.77	**44.55**	24.74	**40.51**

Given Tables 5, 6, and 7, we can observe that a dynamic k significantly increases F1 scores, in most cases within a margin of approximately 0.20% to 30%, offering a balanced trade-off between precision and recall. Although P@1 sometimes yields better results, it only assesses a small part of the models' performance. Employing the F1 score in the evaluation process provides a more accurate picture of the model's performance and balances the precision and recall. Table 8 and Figure 4 demonstrate an example of how dynamic and fixed k work.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Table 8: Examples from the model VM-S trained using FastText monolingual word embeddings on the English-Estonian language pair

Evaluation dataset		**Dynamic *k***	***k***	**Fixed *k* = 1**
legally	juriidiliselt	✓	3	-
	legaalselt	✓		✓
	seaduslikult	✓		-
	õiguslikult	-		-
smoke	suitsu	✓	2	-
	suits	✓		-
	-	-		põleb
bike	rattas	✓	3	-
	ratta	✓		✓
	jalgratas	✓		-

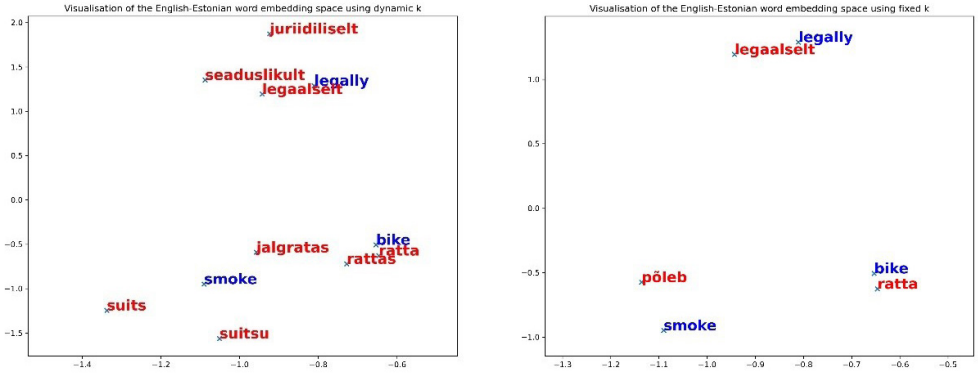

Fig. 4: Visualisation of the English-Estonian word embedding space using dynamic (left) and fixed (right) *k*

When looking at Table 8 and Figure 4, we can see that setting a fixed *k* significantly limits the cross-lingual embedding space by providing only one translation-equivalent candidate with the closest embedding. On the other hand, the classification neural network enables us to set *k* for each headword individually. For example, the headword 'legally' corresponds to four translation equivalents: '*juriidiliselt*', '*legaalselt*', '*seaduslikult*', and '*õiguslikult*'. When using a fixed $k = 1$, the model identified only the closest embedding equivalent, '*legaalselt*'. However, with a dynamic *k*, the model successfully retrieved three of the translation equivalents, i.e., set $k = 3$.

Another interesting example is the incorrectly selected translation-equivalent candidate '*põleb*' by the model using fixed *k*, which did not occur in the evaluation dataset. In the cross-lingual embedding space, '*põleb*' was the closest match. Thus, with $k = 1$ (fixed), the model identified it as the only candidate. Conversely, a dynamic *k* allowed the model to choose candidates whose embeddings were further from the headword's embedding, such as '*suits*' and '*suitsu*' from the evaluation dataset.

6. Conclusion

In this paper, we have introduced cross-lingual embedding models for lexicography, showing the advantages and limitations of certain types of training data, evaluation datasets, and employed metrics. We have demonstrated through extensive analysis across three language pairs that the quality of both training and evaluation data significantly influences the performance of these models. Despite the promising capabilities of models like MUSE, VecMap, and RCLS in bridging linguistic gaps, our findings reveal that enhancements in data quality yield more substantial improvements than algorithmic optimisation alone.

Our investigation into various types of training data underscores the necessity for employing less noisy, more comprehensive monolingual word embeddings such as those from Sketch Engine over commonly used sources like FastText. Similarly, our analysis of evaluation datasets has shown substantial errors and biases that can skew the models' performance. By implementing weighted precision measures that account for the reliability of word pairs in the evaluation datasets, we have provided a more nuanced view of model effectiveness.

Furthermore, the shift from a fixed to a dynamic *k*-based evaluation approach has proven essential in capturing a more accurate representation of model capabilities, as it allows for a balance between precision and recall.

Reformulated, and answering the question posed in the subtitle of our paper, while cross-lingual embedding models are useful tools for computational lexicography, their advancement depends crucially, first, on the quality of the training data, over and above mere algorithmic optimisation. Second, given that not all word pairs in the evaluation datasets have the same reliability, one must and *can* reflect that using weights for the evaluation. Third, given that words often have multiple translation equivalents across languages, one must and *can* also reflect that variation in the model, thereby significantly improving F1 scores.

Future research should continue to focus on refining data quality and developing more robust evaluation methods to fully harness the potential of these models in the field of lexicography. Our work serves as a foundational step towards improving the reliability and efficacy of translation-equivalent retrieval, aiming to support lexicographers in their bilingual dictionary projects with more accurate and resource-efficient tools. Additionally, it is crucial to engage lexicographers in the development process to ensure that these computational advances are practically implementable and truly beneficial in real-world lexicographic practice, addressing the common challenge of bridging theoretical usefulness with practical application.

References

Artetxe, M., Labaka, G., & Agirre, E. (2018a). A Robust Self-learning Method for Fully Unsupervised Cross-lingual Mappings of Word Embeddings. In I. Gurevych, & M. Yusuke

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

(Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 789–798). Association for Computational Linguistics.

Artetxe, M., Labaka, G., & Agirre, E. (2018b). Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 5012–5019. https://doi.org/10.1609/aaai.v32i1.11992_

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & J'egou, H. (2018). Word Translation Without Parallel Data. In *Proceedings of International Conference on Learning Representations* (pp. 1–13).

Denisová, M., & Rychlý, P. (2023). Evaluation of the Cross-lingual Embedding Models from the Lexicographic Perspective. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 1–18). Lexical Computing CZ, s.r.o.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 3483–3487). European Language Resources Association (ELRA).

Herman, O. (2021). Precomputed Embeddings for 15+ Languages. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of the Fifteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2021* (pp. 41–46). Tribun EU.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), P*roceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2979–2984). Association for Computational Linguistics.

Kementchedjhieva, Y., Hartmann, M., & Søgaard, A. (2019). Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. In K. Inui, J, Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3336–3341). Association for Computational Linguistics.

Marchisio, K., Saad-Eldin, A., Duh, K., Carey P., & Koehn, P. (2022). Bilingual Lexicon Induction for Low-Resource Languages using Graph Matching via Optimal Transport. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2545–2561). Association for Computational Linguistics.

Mohiuddin, T., Bari, S. M., & Joty, S. (2020). LNMap: Departures from Isomorphic Assumption in Bilingual Lexicon Induction Through Non-Linear Mapping in Latent Space. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2712–2723). Association for Computational Linguistics.

Ren, S., Liu, S., Zhou, M., & Ma, S. (2020). A Graph-based Coarse-to-fine Method for Unsupervised Bilingual Lexicon Induction. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault

(Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3476–3485). Association for Computational Linguistics.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *The Journal of Artificial Intelligence Research*, *65*, 569–631. https://doi.org/10.1613/jair.1.11640

Sannigrahi, S., & Read, J. (2022). Isomorphic Cross-lingual Embeddings for Low-Resource Languages. In *Proceedings of the 7th Workshop on Representation Learning for NLP (RepL4NLP)* (pp. 133–142). Association for Computational Linguistics.

Smith, N. A. (2019). Contextual Word Representations: A Contextual Introduction. ArXiv, abs/1902.06006.

Vulić, I., Korhonen, A., & Glavaš, G. (2020). Improving Bilingual Lexicon Induction with Unsupervised Post-Processing of Monolingual Word Vector Spaces. In S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, & H. Hajishirzi (Eds.), *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 45–54). Association for Computational Linguistics.

Websites:

Sketch Engine Embeddings. Retrieved May 5, 2024, from https://embeddings.sketchengine.eu/

FastText. Retrieved May 5, 2024, from https://fasttext.cc/

Wikipedia. Retrieved May 5, 2024, from https://www.wikipedia.org/

Contact information

Michaela Denisová
Natural Language Processing Centre, Faculty of Informatics, Masaryk University
449884@mail.muni.cz

Gilles-Maurice de Schryver
Ghent University (BantUGent & LT^3) & University of Pretoria (African Languages)
gillesmaurice.deschryver@UGent.be

Pavel Rychlý
Natural Language Processing Centre, Faculty of Informatics, Masaryk University & Lexical Computing
pary@fi.muni.cz

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.