Small-Scale Cross-Language Authorship Attribution on Social Media Comments

Benjamin Murauer b.murauer@posteo.de
Günther Specht guenther.specht@uibk.ac.at

Abstract

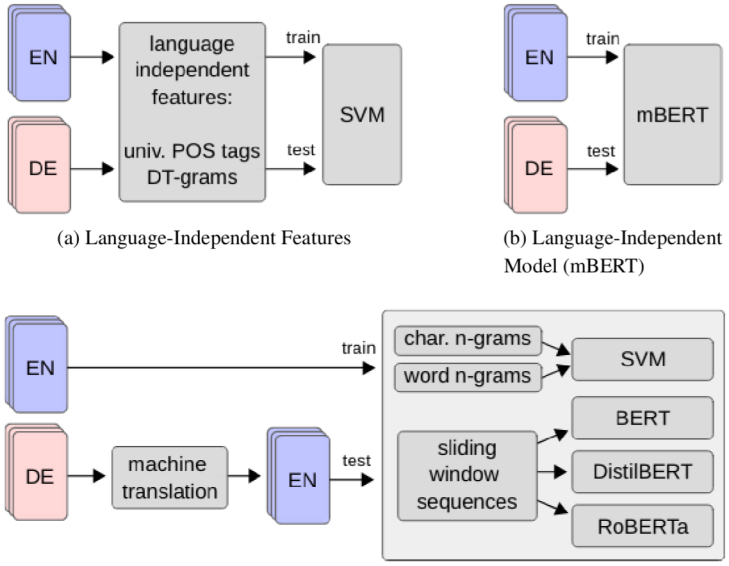
Cross-language authorship attribution is the challenging task of classifying documents by bilingual authors where the training documents are written in a different language than the evaluation documents. Traditional solutions rely on either translation to enable the use of single-language features, or language-independent feature extraction methods. More recently, transformer-based language models like BERT can also be pre-trained on multiple languages, making them intuitive candidates for cross-language classifiers which have not been used for this task yet. We perform extensive experiments to benchmark the performance of three different approaches to a small-scale cross-language authorship attribution experiment: (1) using language-independent features with traditional classification models, (2) using multilingual pre-trained language models, and (3) using machine translation to allow single-language classification. For the language-independent features, we utilize universal syntactic features like part-of-speech tags and dependency graphs, and multilingual BERT as a pre-trained language model. We use a small-scale social media comments dataset, closely reflecting practical scenarios. We show that applying machine translation drastically increases the performance of almost all approaches, and that the syntactic features in combination with the translation step achieve the best overall classification performance. In particular, we demonstrate that pre-trained language models are outperformed by traditional models in small scale authorship attribution problems for every language combination analyzed in this paper.

1 Introduction

In cross-language authorship attribution, the true author of a previously unseen document must be determined from a set of candidate authors after training a machine learning model with documents from those candidates in a different language. Applications for this research include plagiarism detection or other forensic analyses, where the authorship of an incriminating document must be determined, but ground truth texts for comparison of selected suspects are only available in different languages.

The language gap imposes difficulties on the machine learning setup, as the training and testing documents have fewer common features. For example, while some languages may share common words, others use completely different alphabets or writing systems. Therefore, this problem requires one of three general strategies to solve: (1) use machine learning features that don't depend on language, (2) use a model that is inherently capable of solving multilingual problems, or (3) transform one feature space into the other to enable the use of language-dependent features (e.g., by using machine translation).

In this paper, we explore these three approaches for the case of cross-language authorship attribution, depicted in Figure 1.



(a) Language-Independent Features

(b) Language-Independent Model (mBERT)

(c) Translation

Figure 1: The three approaches for cross-language classification models tested in this paper. DE represents any of the other languages from the datasets listed in Table 1. Note that in the experiments, both directions of training and testing are executed (i.e., each approach is also evaluated by training on DE and testing on EN for the depicted dataset).

For the first approach, we use universal part-of-speech (POS) tag n-grams (Nivre et al., 2016) as well as DT-grams (Murauer and Specht, 2021) as language-independent features, paired with a support vector machine as a classifier. For the second approach, we utilize a pre-trained multilingual BERT model, which doesn't require a separate translation pre-processing step, and fine-tune it using training part of the attribution problem. To the best of our knowledge, our study is the first to analyze the performance of this type of model to cross-language authorship attribution problems, representing our first contribution. Finally, we utilize the publicly available Marian NMT machine translation system (Junczys-Dowmunt et al., 2018), and perform several experiments with well-established single-language authorship attribution models, including character n-grams in combination with support vector machines, and also more recent approaches including BERT or RoBERTa.

For these experiments, we make use of datasets consisting of Reddit social media comments compiled by Murauer and Specht (2021). We select bilingual[1] authors that write documents in English as well as one of German, Spanish, French, Dutch and Arabic. By using these small-scale datasets, we provide valid and realistic scenarios for forensic application, are able to skip additional human translation steps required in previously used translation-based datasets, and generalize approaches from previous studies by applying them on a different type of texts and authors. This represents our second contribution.

To ensure the reproducibility of our results and promote future research, all of our code is published online [2].

[1] In the context of this paper, we denote an author writing documents in two languages as bilingual, irrespective of whether both languages are spoken natively by that author, or whether that author was raised bilingually.

[2] `https://git.uibk.ac.at/csak8736/small-scale-authorship-attribution`

2 Related Work

Cross-language text classification problems require different strategies to solve. Some problems allow the usage of parallel corpora for training the model, which enables straightforward transformations of output classes from one language to another (Rasooli et al., 2018). However, parallel corpora are not available for many language combinations, and not suitable for many tasks where the output classes can't be mapped between languages directly. Similarly, other approaches include creating a shared low-dimensional embedding space across languages in an unsupervised pre-processing step (Vulić and Moens, 2015; Mogadala and Rettinger, 2016), but it has been shown that these approaches usually can be outperformed by adding a small amount of supervised cross-language training data (Vulić et al., 2019; Karamanolakis et al., 2020).

Transformer-based pre-trained language models have provided many state-of-the-art results in natural language processing (NLP) in general, and models pre-trained on multiple languages show promising performances in a wide variety of NLP tasks (Devlin et al., 2018; Wu and Dredze, 2019), specifically also in document classification (Wu and Dredze, 2019; Keung et al., 2019). However, to the best of our knowledge, these models have not yet been tested on cross-language authorship attribution problems.

When focussing on authorship attribution, few cross-language studies remain. Llorens and Delany (2016) use differently sized windows in which vocabulary richness measurements are aggregated, requiring very large documents. Bogdanova and Lazaridou (2014) use a variety of different features including the frequency of universal POS tags on attribution, and also utilize machine-translation followed by traditional attribution techniques, providing their best results. However, the dataset that they use consists of translated documents in a single language pair (Spanish - English). More recently, Murauer and Specht (2021) have shown that classifying bilingual authors of social media comments by using universal (language-independent) POS tags can be improved by including dependency grammar information.

In this study, we take inspiration from the latter two studies and compare the performance of grammar-based features to translation-based approaches as well as multilingual language models, which have not been applied to this task. In contrast to similar efforts by Bogdanova and Lazaridou (2014), who classify novels by professional authors, we use small-scale datasets consisting of social media comments, which provide untranslated data and focus on a different text and author type.

3 Datasets

Authorship attribution is the task of determining the authorship of an unknown document given a set of candidate authors. An important difference to other text classification problems is that obtaining more data from a specific target (author) is often not possible. A cross-language setup requires multiple documents from multiple bilingual authors who write in the same two languages, further increasing the difficulty of obtaining large quantities of data.

For this reason, some previous studies have used translated corpora as an alternative means (Llorens and Delany, 2016; Bogdanova and Lazaridou, 2014) to solve the availability issue. There, the author wrote all novels in one language, and translated versions of some of them are used as a source of a different language. While this translation does not fully obfuscate the original authorship (Venuti, 2008), it represents a different scenario as the original authors of those novels did not write them in more than one language. Instead, we want to focus on cross-language features originating from the same author and hence use the corpus presented by Murauer and Specht (2021), which consists of comments form the Reddit social media platform, written by bilingual authors. Here, no additional translation step lies between the originally written documents and the classification model. We further add Arabic as a language from a different group of languages to increase the linguistic diversity.

Languages	Authors	Documents	Avg. Doc. Length	Min. Docs/Auth	Avg. Docs/Auth
EN, DE	10	3,479	3,027	22_{EN}, 20_{DE}	139_{EN}, 69_{DE}
EN, ES	20	4,450	3,125	20_{EN}, 21_{ES}	117_{EN}, 52_{ES}
EN, NL	11	2,410	3,232	20_{EN}, 20_{NL}	154_{EN}, 32_{NL}
EN, FR	45	10,131	3,089	21_{EN}, 20_{FR}	102_{EN}, 61_{FR}
EN, AR	10	2,838	2,117	10_{EN}, 11_{AR}	247_{EN}, 18_{Ar}

Table 1: Datasets used in this paper. The document length is measured in characters.

Table 1 shows the datasets used in this study. In the table, each row represents a dataset consisting of bilingual authors that have written documents in the languages displayed in the first column.

The size of any classification dataset can be divided into two parts; the number of target classes (authors) and the number of training samples (documents) per target class. As both of these numbers differ significantly across the datasets in Table 1, we apply two selection steps before each experiment.

To address the first imbalance and make the results across different language combination directly comparable, we select 10 random authors from each dataset. This number is difficult to increase as it is hard to find bilingual authors in general (e.g., the languages in Table 1 can hardly be considered low-resource by themselves, but the additional restraint on authors writing in multiple languages make even those languages difficult to obtain). On the other hand, it does not influence the evaluation results directly: a dataset with more authors does not automatically imply higher quality of the results, but rather is able to model different scenarios.

Regarding the second imbalance, we select 10 random documents from each author for training, and repeat all experiments five times (each time, choosing 10 random documents) to accommodate for this imbalance. This way, each author receives the same number of training documents. We choose 10 as this is the lower bound of how many documents an author has written (in the Arabic dataset). Note that we do not restrict the number of documents used for testing, as it does not influence the training process of the machine learning models, but rather helps to increase the confidence of the evaluation results.

We want to emphasize that having few authors and few documents per author is a valid and realistic scenario for many applications, and therefore, the small size of the datasets is a challenging and central corner stone of our work, rather than a limitation.

4 Methodology

We test three different approaches for cross-language attribution, which are depicted in Figure 1. We follow the same evaluation strategy with each approach: For each dataset, we perform all experiments in two directions, (1) train with the English part of the dataset and test with the respective other language, and (2) the other way around.

Each subsection will discuss any (hyper)parameters of the respective model, the full list of these parameters is shown in Table 2 as a reference.

4.1 Language-Independent Features

In this work, we make use of two language-independent features based on syntactic information, as this type of features has been successfully been used in previous studies (Bogdanova and Lazaridou, 2014; Tschuggnall and Specht, 2014).

Hyperparameters (used in grid search)	
character, word, universal POS tag n-gram size	$n \in [1-3]$
DT-gram shape[1]	$\mathrm{DT}_{\mathrm{anc}}$, $\mathrm{DT}_{\mathrm{sib}}$, $\mathrm{DT}_{\mathrm{pq}}$, $\mathrm{DT}_{\mathrm{inv}}$
DT-gram parameter sizes[1]	$\mathrm{sib}, \mathrm{anc} \in [1-4]$
support vector machine regularization factor C	$C \in [0.1, 1, 10]$
Language model parameters (static)	
Fine-tuning epochs	3
Max. sequence length	256
Learning rate	4×10^{-5}
Batch size	8

Table 2: Parameters used in the models. [1]Parameters of the DT-grams features by Murauer and Specht (2021).

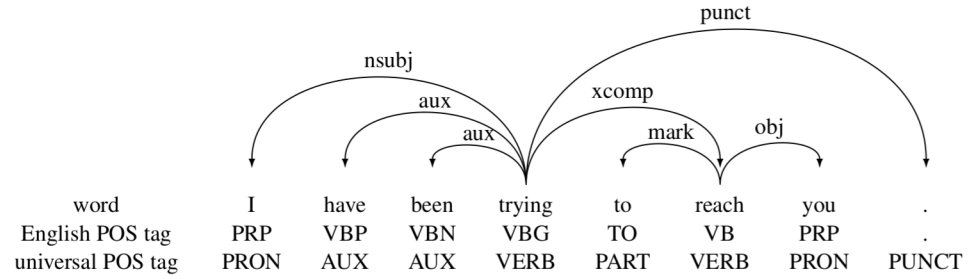


Figure 2: Differences between English-specific and universal POS tags of the sentence 'I have been trying to reach you.'.

1. universal POS tags are the result of a universal mapping of (normally language-dependent) POS tags to a language-independent space, called *universal* POS tags (Nivre et al., 2016). Figure 2 shows both the English-specific POS tags as well as their universal mappings for each word of the sentence "I have been trying to reach you". It can be seen that the mapping produces coarser relationships (e.g., the information that "trying" is a present participle is lost) but enables direct comparisons of POS tags across different languages. From the resulting POS tags, we construct n-grams by using each tag as a token.

2. DT-grams are dependency graph substructures introduced by Murauer and Specht (2021). In addition to using POS tags, the relationship between words is captured. Similar to the POS tags themselves, these dependencies can also be mapped to a language-independent space using universal dependencies (Nivre et al., 2017). We choose the same substructure layout candidates that the original authors suggest, and perform a grid search to determine the optimal candidate as well as the optimal values for the two parameters that each substructure has, from a range of $[1-4]$ (cf. Table 2).

We use a linear support vector machine as a classifier for both approaches, which has been shown to be an effective model for authorship attribution (Stamatatos, 2013; Tschuggnall et al., 2019). We utilize the *stanza* library (Qi et al., 2020) to obtain both the universal POS tags as well as the universal dependency graphs for each sentence in each dataset.

4.2 Pre-Trained Multi-Language Models

We use the multilingual version of BERT called mBERT (Devlin et al., 2018), which is pre-trained using 100 languages and has been successfully applied in many different cross-language text classification tasks (Wu and Dredze, 2019; Keung et al., 2019). While other pre-trained models in multiple languages exist, none of them cover all languages presented in this paper. We use the parameters suggested by the original authors, which are listed in Table 2. As all transformer-based models, mBERT operates on sequences of words, and the maximal length of these sequences is determined by the pre-training step of the model (which is 768 tokens for mBERT). Since the documents used in our classification setup are significantly larger than this limit, we use a sliding window approach to generate multiple samples from each document, so that every part of each document is used for fine-tuning. Thereby, each window overlaps 20% with the previous one.

The results of this model are especially useful to answer the question of whether it is more effective to translate documents in order to be able to use single-language classification models, or if inherently multilingual models are able to render this additional step superfluous.

4.3 Translation

We use the Marian NMT machine translation models (Junczys-Dowmunt et al., 2018) which are available for many language combinations. While the library offers models for both directions, for each dataset, we translate the non-English documents to English rather than the other way around, as there are more pre-trained language models available for English, and the multilingual version of BERT is also pre-trained with more English data. We therefore have more opportunities to compare to other single-language models, and expect mBERT to perform better.

At this point, we test different classification approaches on the now single-language dataset. As suggested in previous (mono-lingual) authorship research (Stamatatos, 2013; Tschuggnall and Specht, 2014), we use a linear support vector machine in combination with frequencies of character 3-grams and word unigrams. The hyperparameters of these models are listed in Table 2. We also analyze the syntactic features from approach 1, but skip the mapping of the POS tags to the universal space. This way, all POS tags are English-specific and therefore finer-grained, increasing the vocabulary size of these features.

Our experiments further include three mono-lingual pre-trained language models: mono-lingual BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019). The parameters of these models are set according to recommendations of the original authors, and are listed in Table 2. Following our approach for mBERT, we apply the sliding window scheme to generate samples that fit in the respective maximal sequence lengths.

5 Results

Table 3 shows the results of the three presented approaches for all datasets, where the score is measured in macro-averaged F1.

The results of the language-independent features show that while adding dependency information to the universal POS tag features increases the F1 score for all language pairs consistently, the extent of this increase differs and is most clearly visible for the English/German dataset.

The language-independent model mBERT outperforms the DT-grams+SVM model for some language combinations, but not for the English/German and English/Dutch dataset.

The DT-grams exclusively capture grammatical features while the mBERT model incorporates content-based properties, making the two feature categories largely independent from each other. We therefore suspect them to be suitable candidates for ensemble models, which we aim to pursue in future work.

	ar	de	es	fr	nl
Approach 1: Language-independent models on untranslated documents					
Universal POS tag n-grams	0.110	0.260	0.290	0.239	0.226
DT-grams	0.175	0.400	0.317	0.283	0.262
Approach 2: Multilingual pre-trained language model					
Multilingual BERT	0.228	0.242	0.382	0.368	0.250
Approach 3: Single language models on translated documents					
Character n-grams + SVM	0.375	0.410	0.443	0.443	0.398
Word n-grams + SVM	**0.380**	0.360	0.428	0.413	0.390
BERT	0.291	0.273	0.342	0.425	0.308
DistilBERT	0.160	0.157	0.194	0.186	0.160
RoBERTa	0.298	0.261	0.382	0.432	0.311
English POS tag n-grams	0.281	0.465	0.455	0.503	0.419
English DT-grams	0.347	**0.467**	**0.465**	**0.552**	**0.433**
Combined: Language-independent models on translated documents					
Universal POS tag n-grams	0.256	0.322	0.388	0.362	0.354
DT-grams	0.327	0.435	0.456	0.447	0.416
Multilingual BERT	0.286	0.273	0.322	0.425	0.308

Table 3: Classification score measured in $F1_{macro}$ of the three different approaches, as well as a combination where all language-independent models are applied to the translated documents.

Both the grammar-based features in combination with the support vector machine as well as the multilingual BERT model are outperformed by the translation approach using the character- and word-based n-grams. The former confirms the results of Bogdanova and Lazaridou (2014), while the latter is a novel result showing that a multilingual BERT model is less efficient for such small datasets. In total, the English-specific syntactic features of the machine-translated documents show the best average performance consistently across many different languages. Only for the Arabic dataset, the word n-grams produce the best results.

Machine-translation also improves the performance of the models using language-independent features. The relevant part of our results in this regard is visible in the *Combined* part of Table 3 and shows that translation is able to boost the F1 scores of almost all languages and models, except for the Spanish dataset in combination with mBERT. These results suggest that previous findings by Bogdanova and Lazaridou (2014) are not restricted to professionally written novels, but also apply to small social media datasets. Moreover, the differences between the universal and English-specific grammar-based features demonstrate that the reduced POS tag vocabulary allowing cross-language analyses comes with a notable performance loss.

While the multilingual BERT model is able to compete with the other language-independent features, it's performance is well below all methods using machine-translation. In general, all language models are outperformed by syntactic and lexicographic features in the respective approaches, signaling that the datasets are too small for fine-tuning them sufficiently. We observe an amplification of this effect on DistilBERT, which suggests that models produced by knowledge distillation are more susceptible to smaller datasets than their original teacher model.

Summarized, our findings suggest that for cross-language authorship attribution at a small scale, machine-translation is a highly efficient first step in every case, and syntactic features are a promising candidate for datasets of this size.

6 Limitations

In general, the different datasets show a varying performance, where documents in some languages (e.g., German) are easier to attribute than others (e.g., Arabic). We attribute this effect to the linguistic distance between the language pairs (i.e., German and English are closer related to each other than Arabic and English). More datasets containing additional language pairs are required for more comprehensive comparisons in this regard.

By design, the results using machine translation in the fashion presented in this paper depend on the quality of these translation models. Especially for low-resource languages, this means that differences in translation quality between different language pairs are likely to influence the final attribution results.

7 Conclusion

In this paper, we have demonstrated different approaches to the problem of cross-language authorship attribution for bilingual authors writing in both English and one of Arabic, German, Spanish, French, and Dutch. We have analyzed language-independent syntactic features, using multilingual pre-trained language models as well as performing machine translation followed by several single-language solutions. Eventually, we show that for small-scale problems with very few training documents, using machine translation followed by models using lexicographic and syntactic features yields the best results for all languages analyzed in this work.

In the near future, we want to focus on the influence of the dataset size on the pre-trained language models to see how much data is required for these models to succeed in authorship attribution tasks. Also, we want to investigate recent work suggesting that small translation dictionaries represent a suitable substitution for full translation (Karamanolakis et al., 2020), which is a time and resource-consuming process.

References

Bogdanova, D. and Lazaridou, A. (2014). Cross-language authorship attribution. In *Ninth International Conference on Language Resources and Evaluation (LREC'2014)*, pages 2015–2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Karamanolakis, G., Hsu, D., and Gravano, L. (2020). Cross-lingual text classification with minimal resources by transferring a sparse teacher. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3604–3622.

Keung, P., Lu, Y., and Bhardwaj, V. (2019). Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner. http://arxiv.org/pdf/1909.00153v3.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. http://arxiv.org/pdf/1907.11692v1.

Llorens, M. and Delany, S. J. (2016). Deep level lexical features for cross-lingual authorship attribution. In *Proceedings of the first Workshop on Modeling, Learning and Mining for Cross/Multilinguality*, pages 16–25. Dublin Institute of Technology.

Mogadala, A. and Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702.

Murauer, B. and Specht, G. (2021). DT-grams: Structured Dependency Grammar Stylometry for Cross-Language Authorship Attribution. `https://arxiv.org/pdf/2106.05677`.

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., et al. (2017). Universal dependencies 2.1. `https://universaldependencies.org/`.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th Int. Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Rasooli, M. S., Farra, N., Radeva, A., Yu, T., and McKeown, K. (2018). Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1):143–165.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. `https://arxiv.org/pdf/1910.01108`.

Stamatatos, E. (2013). On the Robustness of Authorship Attribution Based on Character N-Gram Features. *Journal of Law & Policy*, pages 421–439.

Tschuggnall, M., Murauer, B., and Specht, G. (2019). Reduce & attribute: Two-step authorship attribution for large-scale problems. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 951–960, Hong Kong, China. Association for Computational Linguistics.

Tschuggnall, M. and Specht, G. (2014). Enhancing Authorship Attribution By Utilizing Syntax Tree Profiles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'2014)*, volume 2, pages 195–199. Association for Computational Linguistics.

Venuti, L. (2008). *The translator's invisibility: A history of translation*. Routledge.

Vulić, I., Glavaš, G., Reichart, R., and Korhonen, A. (2019). Do we really need fully unsupervised cross-lingual embeddings? `https://arxiv.org/pdf/1909.01638`.

Vulić, I. and Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International Conference on Research and Development in Information Retrieval*. ACM Press.

Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. `https://arxiv.org/pdf/1904.09077v2`.