Why Overfitting Isn't Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries

Mozhi Zhang*
CS and UMIACS
University of Maryland
mozhi@cs.umd.edu

Yoshinari Fujinuma*
Computer Science
University of Colorado
fujinumay@gmail.com

Michael J. Paul
Information Science
University of Colorado
mpaul@colorado.edu

Jordan Boyd-Graber
UMD CS, iSchool, UMIACS, and LSC
and Google Research Zürich
jbg@boydgraber.org

Abstract

Cross-lingual word embeddings (CLWE) are often evaluated on bilingual lexicon induction (BLI). Recent CLWE methods use linear projections, which underfit the training dictionary, to generalize on BLI. However, underfitting can hinder generalization to other downstream tasks that rely on words from the training dictionary. We address this limitation by *retrofitting* CLWE to the training dictionary, which pulls training translation pairs closer in the embedding space and overfits the training dictionary. This simple post-processing step often improves accuracy on two downstream tasks, despite lowering BLI test accuracy. We also retrofit to both the training dictionary and a synthetic dictionary induced from CLWE, which sometimes generalizes even better on downstream tasks. Our results confirm the importance of fully exploiting the training dictionary in downstream tasks and explains why BLI is a flawed CLWE evaluation.

1 Introduction

Cross-lingual word embeddings (CLWE) map words across languages to a shared vector space. Recent supervised CLWE methods follow a projection-based pipeline (Mikolov et al., 2013). Using a training dictionary, a linear projection maps pre-trained monolingual embeddings to a multilingual space. While CLWE enable many multilingual tasks (Klementiev et al., 2012; Guo et al., 2015; Zhang et al., 2016; Ni et al., 2017), most recent work only evaluates CLWE on bilingual lexicon induction (BLI). Specifically, a set of test words are translated with a retrieval heuristic (e.g., nearest neighbor search) and compared against gold translations. BLI accuracy is easy to compute and captures the desired property of CLWE that translation pairs should be close. However, BLI accuracy does not always correlate with accuracy on downstream tasks such as cross-lingual document classification and dependency parsing (Ammar et al., 2016; Fujinuma et al., 2019; Glavas et al., 2019).

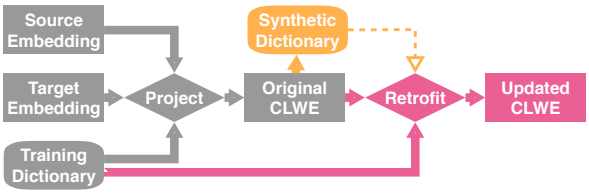

Figure 1: To fully exploit the training dictionary, we retrofit projection-based CLWE to the training dictionary as a post-processing step (pink parts). To preserve correctly aligned translations in the original CLWE, we optionally retrofit to a synthetic dictionary induced from the original CLWE (orange parts).

Let's think about why that might be. BLI accuracy is only computed on *test* words. Consequently, BLI hides linear projection's inability to align all training translation pairs at once; i.e., projection-based CLWE *underfit* the training dictionary. Underfitting does not hurt BLI test accuracy, because test words are excluded from the training dictionary in BLI benchmarks. However, words from the training dictionary may be nonetheless predictive in downstream tasks; e.g., if "good" is in the training dictionary, knowing its translation is useful for multilingual sentiment analysis.

In contrast, *overfitting* the training dictionary hurts BLI but can improve downstream models. We show this by adding a simple post-processing step to projection-based pipelines (Figure 1). After training supervised CLWE with a projection, we *retrofit* (Faruqui et al., 2015) the CLWE to the same training dictionary. This step pulls training translation pairs closer and overfits: the updated embeddings have perfect BLI training accuracy, but

*Equal contribution

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2214–2220
July 5 - 10, 2020.

BLI *test* accuracy drops. Empirically, retrofitting improves accuracy in two downstream tasks other than BLI, confirming the importance of fully exploiting the training dictionary.

Unfortunately, retrofitting to the training dictionary may inadvertently push some translation pairs further away. To balance between fitting the training dictionary and generalizing on other words, we explore retrofitting to both the training dictionary and a *synthetic dictionary* induced from the CLWE. Adding the synthetic dictionary keeps some correctly aligned translations in the original CLWE and can further improve downstream models by striking a balance between training and test BLI accuracy.

In summary, our contributions are two-fold. First, we explain *why* BLI does not reflect downstream task accuracy. Second, we introduce two post-processing methods to improve downstream models by fitting the training dictionary better.

2 Limitation of Projection-Based CLWE

This section reviews projection-based CLWE. We then discuss how BLI evaluation obscures the limitation of projection-based methods.

Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ be a pre-trained d-dimensional word embedding matrix for a source language, where each column $\mathbf{x}_i \in \mathbb{R}^d$ is the vector for word i from the source language with vocabulary size n, and let $\mathbf{Z} \in \mathbb{R}^{d \times m}$ be a pre-trained word embedding matrix for a target language with vocabulary size m. Projection-based CLWE maps $\mathbf{X}$ and $\mathbf{Z}$ to a shared space. We focus on supervised methods that learn the projection from a training dictionary $\mathcal{D}$ with translation pairs (i, j).

Mikolov et al. (2013) first propose projection-based CLWE. They learn a linear projection $\mathbf{W} \in \mathbb{R}^{d \times d}$ from $\mathbf{X}$ to $\mathbf{Z}$ by minimizing distances between translation pairs in a training dictionary:

$$\min_{\mathbf{W}} \sum_{(i,j) \in \mathcal{D}} \|\mathbf{W}\mathbf{x}_i - \mathbf{z}_j\|_2^2. \quad (1)$$

Recent work improves this method with different optimization objectives (Dinu et al., 2015; Joulin et al., 2018), orthogonal constraints on $\mathbf{W}$ (Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017), pre-processing (Zhang et al., 2019), and subword features (Chaudhary et al., 2018; Czarnowska et al., 2019; Zhang et al., 2020).

Projection-based methods *underfit*—a linear projection has limited expressiveness and cannot perfectly align all training pairs. Unfortunately, this weakness is not transparent when using BLI as the standard evaluation for CLWE, because BLI test sets omit training dictionary words. However, when the training dictionary covers words that help downstream tasks, underfitting limits generalization to other tasks. Some BLI benchmarks use frequent words for training and infrequent words for testing (Mikolov et al., 2013; Conneau et al., 2018). This mismatch often appears in real-world data, because frequent words are easier to find in digital dicitonaries (Czarnowska et al., 2019). Therefore, training dictionary words are often more important in downstream tasks than test words.

3 Retrofitting to Dictionaries

To fully exploit the training dictionary, we explore a simple post-processing step that *overfits* the dictionary: we first train projection-based CLWE and then *retrofit* to the training dictionary (pink parts in Figure 1). Retrofitting was originally introduced for refining monolingual word embeddings with synonym constraints from a lexical ontology (Faruqui et al., 2015). For CLWE, we retrofit using the training dictionary $\mathcal{D}$ as the ontology.

Intuitively, retrofitting pulls translation pairs closer while minimizing deviation from the original CLWE. Let $\mathbf{X}'$ and $\mathbf{Z}'$ be CLWE trained by a projection-based method, where $\mathbf{X}' = \mathbf{W}\mathbf{X}$ are the projected source embeddings and $\mathbf{Z}' = \mathbf{Z}$ are the target embeddings. We learn new CLWE $\hat{\mathbf{X}}$ and $\hat{\mathbf{Z}}$ by minimizing

$$L = L_a + L_b, \quad (2)$$

where L_a is the squared distance between the updated CLWE from the original CLWE:

$$L_a = \alpha\|\hat{\mathbf{X}} - \mathbf{X}'\|^2 + \alpha\|\hat{\mathbf{Z}} - \mathbf{Z}'\|^2, \quad (3)$$

and L_b is the total squared distance between translations in the dictionary:

$$L_b = \sum_{(i,j) \in \mathcal{D}} \beta_{ij}\|\hat{\mathbf{x}}_i - \hat{\mathbf{z}}_j\|^2. \quad (4)$$

We use the same α and β as Faruqui et al. (2015) to balance the two objectives.

Retrofitting tends to overfit. If α is zero, minimizing L_b collapses each training pair to the same vector. Thus, all training pairs are perfectly aligned. In practice, we use a non-zero α for regularization, but the updated CLWE still have perfect training BLI accuracy (Figure 2). If the training dictionary covers predictive words, we expect retrofitting to improve downstream task accuracy.

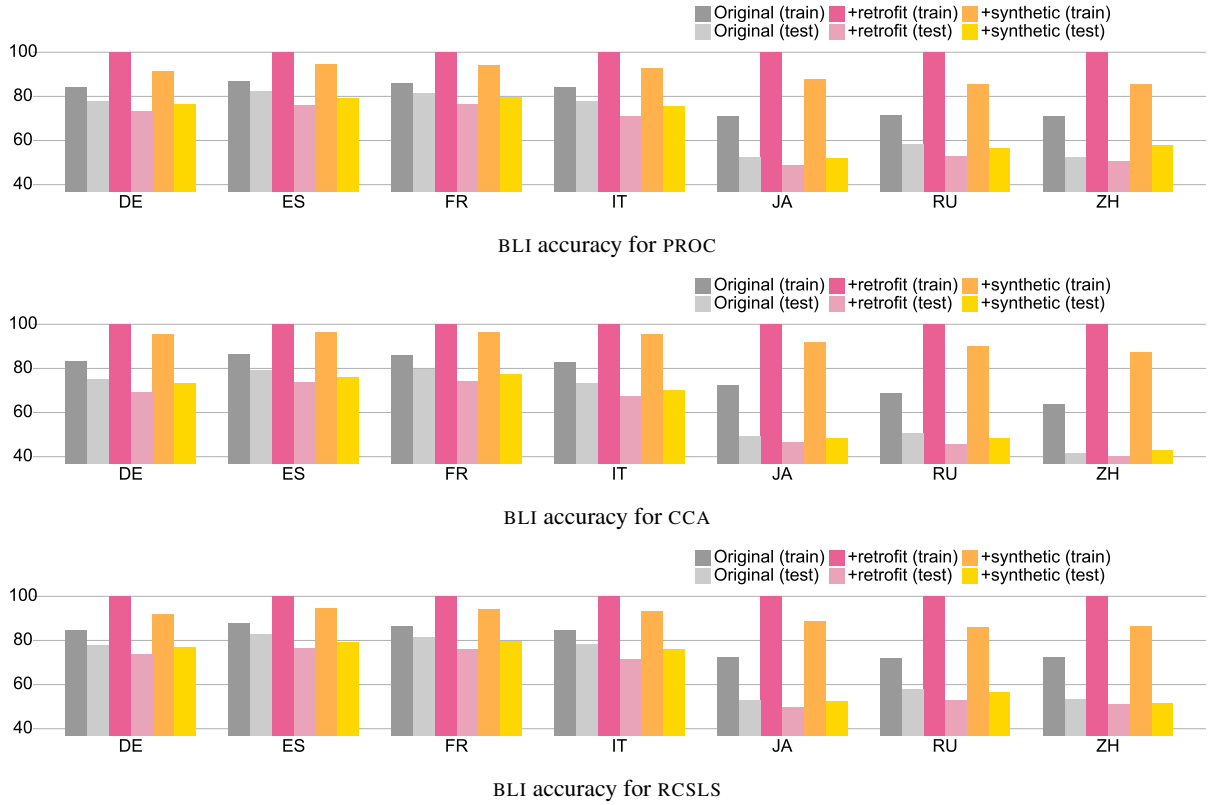

Figure 2: Train and test accuracy (P@1) for BLI on MUSE; Projection-based CLWE underfit the training dictionary (gray), but retrofitting to the training dictionary overfits (pink). Adding a synthetic dictionary balances between training and test accuracy (orange).

3.1 Retrofitting to Synthetic Dictionary

While retrofitting brings pairs in the training dictionary closer, the updates may also separate translation pairs outside of the dictionary because retrofitting ignores words outside the training dictionary. This can hurt both BLI test accuracy and downstream task accuracy. In contrast, projection-based methods underfit but can discover translation pairs outside the training dictionary. To keep the original CLWE's correct translations, we retrofit to both the training dictionary and a *synthetic dictionary* induced from CLWE (orange, Figure 1).

Early work induces dictionaries from CLWE through nearest-neighbor search (Mikolov et al., 2013). We instead use cross-domain similarity local scaling (Conneau et al., 2018, CSLS), a translation heuristic more robust to hubs (Dinu et al., 2015) (a word is the nearest neighbor of many words). We build a synthetic dictionary $\mathcal{D}'$ with word pairs that are *mutual* CSLS nearest neighbors. We then retrofit the CLWE to a combined dictionary $\mathcal{D} \cup \mathcal{D}'$. The synthetic dictionary keeps closely aligned word pairs in the original CLWE, which sometimes improves downstream models.

4 Experiments

We retrofit three projection-based CLWE to their training dictionaries and synthetic dictionaries.[1] We evaluate on BLI and two downstream tasks. While retrofitting decreases test BLI accuracy, it often improves downstream models.

4.1 Embeddings and Dictionaries

We align English embeddings with six target languages: German (DE), Spanish (ES), French (FR), Italian (IT), Japanese (JA), and Chinese (ZH). We use 300-dimensional fastText vectors trained on Wikipedia and Common Crawl (Grave et al., 2018). We lowercase all words, only keep the 200K most frequent words, and apply five rounds of Iterative Normalization (Zhang et al., 2019).

We use dictionaries from MUSE (Conneau et al., 2018), a popular BLI benchmark, with standard splits: train on 5K source word translations and test on 1.5K words for BLI. For each language, we train three projection-based CLWE: canonical correlation analysis (Faruqui and Dyer, 2014, CCA),

[1] Code at https://go.umd.edu/retro_clwe.

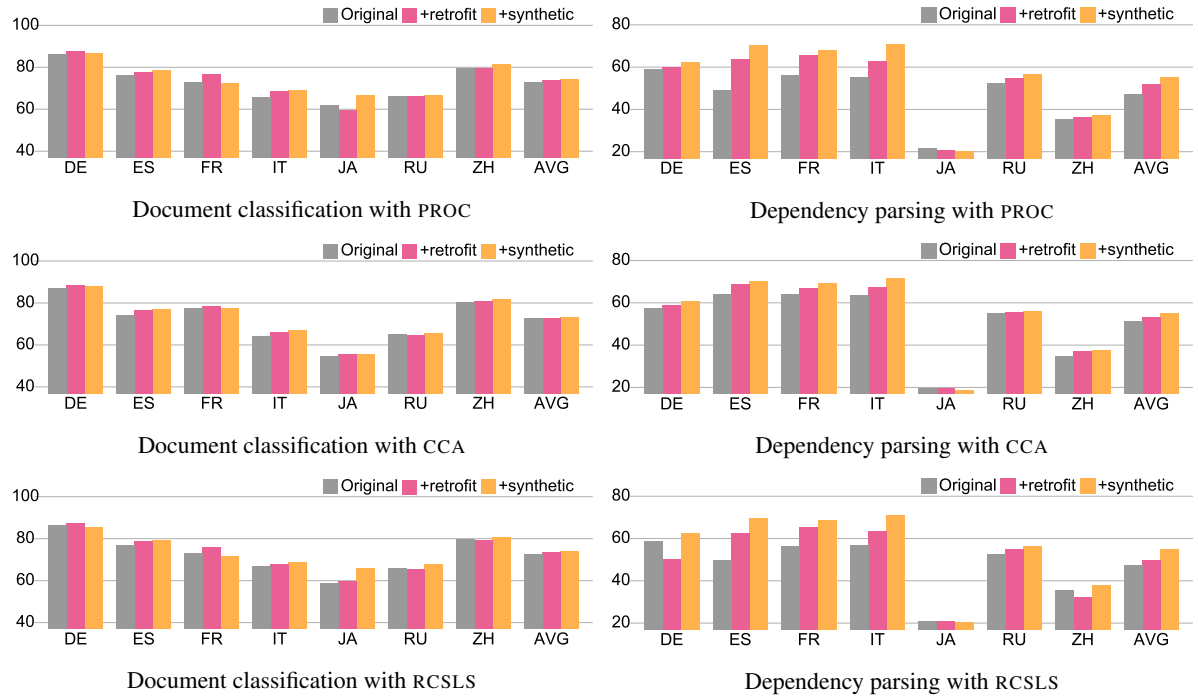

Figure 3: For each CLWE, we report accuracy for document classification (left) and unlabeled attachment score (UAS) for dependency parsing (right). Compared to the original embeddings (gray), retrofitting to the training dictionary (pink) improves average downstream task scores, confirming that fully exploiting the training dictionary helps downstream tasks. Adding a synthetic dictionary (orange) further improves test accuracy in some languages.

Procrustes analysis (Conneau et al., 2018, PROC), and Relaxed CSLS loss (Joulin et al., 2018, RCSLS). We retrofit these CLWE to the training dictionary (pink in figures) and to both the training and the synthetic dictionary (orange in figures).

In MUSE, words from the training dictionary have higher frequencies than words from the test set.[2] For example, the most frequent word in the English-French test dictionary is "torpedo", while the training dictionary has translations for frequent words such as "the" and "good". As discussed in §2, more frequent words are likely to be more salient in downstream tasks, so underfitting these more frequent training pairs hurts generalization to downstream tasks.[3]

4.2 Intrinsic Evaluation: BLI

We first compare BLI accuracy on both training and test dictionaries (Figure 2). We use CSLS to translate words with default parameters. The original projection-based CLWE have the highest test accuracy but underfit the training dictionary. Retrofitting to the training dictionary perfectly fits the training dictionary but drops test accuracy. Retrofitting to the combined dictionary splits the difference: higher test accuracy but lower train accuracy. These three modes offer a continuum between BLI test and training accuracy.

4.3 Extrinsic Evaluation: Downstream Tasks

We compare CLWE on two downstream tasks: document classification and dependency parsing. We fix the embeddng layer of the model to CLWE and use the zero-shot setting, where a model is trained in English and evaluated in the target language.

Document Classification Our first downstream task is document-level classification. We use MLDoc, a multilingual classification benchmark (Schwenk and Li, 2018) using the standard split with 1K training and 4K test documents. Following Glavas et al. (2019), we use a convolutional neural network (Kim, 2014). We apply 0.5 dropout to the final layer, run Adam (Kingma and Ba, 2015) with default parameters for ten epochs, and report the average accuracy of ten runs.

Dependency Parsing We also test on dependency parsing, a structured prediction task. We use Universal Dependencies (Nivre et al., 2019,

[2] https://github.com/facebookresearch/MUSE/issues/24

[3] A pilot study confirms that retrofitting to infrequent word pairs is less effective.

v2.4) with the standard split. We use the biaffine parser (Dozat and Manning, 2017) in AllenNLP (Gardner et al., 2017) with the same hyperparameters as Ahmad et al. (2019). To focus on the influence of CLWE, we remove part-of-speech features (Ammar et al., 2016). We report the average unlabeled attachment score (UAS) of five runs.

Results Although training dictionary retrofitting lowers BLI test accuracy, it improves both downstream tasks' test accuracy (Figure 3). This confirms that over-optimizing the test BLI accuracy can hurt downstream tasks because training dictionary words are also important. The synthetic dictionary further improves downstream models, showing that generalization to downstream tasks must balance between BLI training and test accuracy.

Qualitative Analysis As a qualitative example, coordinations improve after retrofitting to the training dictionary. For example, in the German sentence "Das Lokal ist sauber, hat einen gemütlichen 'Raucherraum' und wird gut besucht", the bar ("Das Lokal") has three properties: it is clean, has a smoking room, and is popular. However, without retrofitting, the final property "besucht" is connected to "hat" instead of "sauber"; i.e., the final clause stands on its own. After retrofitting to the English-German training dictionary, "besucht" is moved closer to its English translation "visited" and is correctly parsed as a property of the bar.

5 Related Work

Previous work proposes variants of retrofitting broadly called *semantic specialization* methods. Our pilot experiments found similar trends when replacing retrofitting with Counter-fitting (Mrkšić et al., 2016) and Attract-Repel (Mrkšić et al., 2017), so we focus on retrofitting.

Recent work applies semantic specialization to CLWE by using multilingual ontologies (Mrkšić et al., 2017), transferring a monolingual ontology across languages (Ponti et al., 2019), and asking bilingual speakers to annotate task-specific keywords (Yuan et al., 2019). We instead re-use the training dictionary of the CLWE.

Synthetic dictionaries are previously used to iteratively refine a linear projection (Artetxe et al., 2017; Conneau et al., 2018). These methods still underfit because of the linear constraint. We instead retrofit to the synthetic dictionary to fit the training dictionary better while keeping some generalization power of projection-based CLWE.

Recent work investigates cross-lingual contextualized embeddings as an alternative to CLWE (Eisenschlos et al., 2019; Lample and Conneau, 2019; Huang et al., 2019; Wu and Dredze, 2019; Conneau et al., 2020). Our method may be applicable, as recent work also applies projections to contextualized embeddings (Aldarmaki and Diab, 2019; Schuster et al., 2019; Wang et al., 2020; Wu et al., 2020).

6 Conclusion and Discussion

Popular CLWE methods are optimized for BLI test accuracy. They underfit the training dictionary, which hurts downstream models. We use retrofitting to fully exploit the training dictionary. This post-processing step improves downstream task accuracy despite lowering BLI test accuracy. We then add a synthetic dictionary to balance BLI test and training accuracy, which further helps downstream models on average.

BLI test accuracy does not always correlate with downstream task accuracy because words from the training dictionary are ignored. An obvious fix is adding training words to the BLI test set. However, it is unclear how to balance between training and test words. BLI accuracy assumes that all test words are equally important, but the importance of a word depends on the downstream task; e.g., "the" is irrelevant in document classification but important in dependency parsing. Therefore, future work should focus on downstream tasks instead of BLI.

We focus on retrofitting due to its simplicity. There are other ways to fit the dictionary better; e.g., using a non-linear projection such as a neural network. We leave the exploration of non-linear projections to future work.

Acknowledgement

This research is supported by NSF grant IIS-1564275 and by ODNI, IARPA, via the BETTER Program contract #2019-19051600005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of Empirical Methods in Natural Language Processing*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the Association for Computational Linguistics*.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of Empirical Methods in Natural Language Processing*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Association for Computational Linguistics*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations*.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! A comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of Empirical Methods in Natural Language Processing*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the International Conference on Learning Representations Workshop Track*.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the International Conference on Learning Representations*.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of Empirical Methods in Natural Language Processing*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. *Conference of the North American Chapter of the Association for Computational Linguistics*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul. 2019. A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. In *Proceedings of the Association for Computational Linguistics*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the Association for Computational Linguistics*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Language Resources and Evaluation Conference*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the Association for Computational Linguistics*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of Empirical Methods in Natural Language Processing*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of Empirical Methods in Natural Language Processing*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of International Conference on Computational Linguistics*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proceedings of Advances in Neural Information Processing Systems*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the Association for Computational Linguistics*.

Joakim Nivre, Mitchell Abrams, Željko Agić, and et al. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of Empirical Methods in Natural Language Processing*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Language Resources and Evaluation Conference*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the International Conference on Learning Representations*.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *Proceedings of the International Conference on Learning Representations*.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the Association for Computational Linguistics*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of Empirical Methods in Natural Language Processing*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2019. Interactive refinement of cross-lingual word embeddings. *arXiv preprint arXiv:1911.03070*.

Mozhi Zhang, Yoshinari Fujinuma, and Jordan Boyd-Graber. 2020. Exploiting cross-lingual subword similarities in low-resource document classification. In *Association for the Advancement of Artificial Intelligence*.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the Association for Computational Linguistics*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Conference of the North American Chapter of the Association for Computational Linguistics*.