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

Author = {Mozhi Zhang and Yoshinari Fujinuma and Michael J. Paul and Jordan Boyd-Graber},
Title = {Why Overfitting Isn’t Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries},
Booktitle = {Association for Computational Linguistics},
Year = {2020},
Location = {The Cyberverse Simulacrum of Seattle},
Url = {http://umiacs.umd.edu/~jbg/docs/2020_acl_refine.pdf},

Accessible Abstract: Computers need to represent words in a computer-readable way. This work talks about how slightly moving these representations for words in different languages to be closer to a small list of translations (like from a dictionary) after doing fancy machine learning works better on downstream tasks (e.g., guessing grammatical category of a word) but hurts on asking the algorithm for translations of unseen words.

Links:
- Code [https://go.umd.edu/retro_clwe](https://go.umd.edu/retro_clwe)

Downloaded from [http://umiacs.umd.edu/~jbg/docs/2020_acl_refine.pdf](http://umiacs.umd.edu/~jbg/docs/2020_acl_refine.pdf)

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.
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).

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

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

$$
\min_W \sum_{(i,j) \in D} \|WX_i - Z_j\|_2^2.
$$

Recent work improves this method with different optimization objectives (Dinu et al., 2015; Joulin et al., 2018), orthogonal constraints on $W$ (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 dictionaries (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 $D$ as the ontology.

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

$$
L = L_a + L_b,
$$

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

$$
L_a = \alpha \|\hat{X} - X'\|^2 + \alpha \|\hat{Z} - Z'\|^2,
$$

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

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

We use the same $\alpha$ and $\beta$ as Faruqui et al. (2015) to balance the two objectives.

Retrofitting tends to overfit. If $\alpha$ 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 $\alpha$ 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 \( D' \) with word pairs that are \textit{mutual} CSLS nearest neighbors. We then retrofit the CLWE to a combined dictionary \( D \cup D' \). 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.\footnote{Code at https://go.umd.edu/retro_clwe.} 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 lower case 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),
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. 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.  

4.3 Extrinsic Evaluation: Downstream Tasks

We compare CLWE on two downstream tasks: document classification and dependency parsing. We fix the embedding 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,
v2.4) with the standard split. We use the bi-affine parser (Dozat and Manning, 2017) in AllenNLP (Gardner et al., 2017) with the same hyper-parameters 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


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.


