Investigating the Use of BERT Anchors for Bilingual Lexicon Induction with Minimal Supervision

Anonymous ACL submission

Abstract

This paper investigates the use of static anchors from transformer architectures for the task of Bilingual Lexicon Induction. We revisit an existing approach built around the ELMo architecture and explore the use of the methodology on the BERT family of language models. Experiments are performed and analysed for three language pairs, combining English with three target languages from very different language families, Hindi, Dutch, and Russian. Although the contextualised approach is not able to outperform the SOTA VecMap method, we find that it is easily adaptable to newer transformer models and can compete with the MUSE approach. An error analysis reveals interesting trends accross languages and shows how the method could be further improved by building on the basic hypothesis that transformer embeddings can indeed be decomposed into a static anchor and a dynamic context component. We make the code, the extracted anchors before and after alignement and the modified train and test sets available for use.[1]

1 Introduction

Despite the great progress witnessed in recent years for various NLP tasks, low(er)-resourced languages are often lagging behind because of data scarcity. To overcome this lack of resources, researchers have started to investigate the use of cross-lingual information, where knowledge or data from a rich-resourced language, like English, is used to improve the modeling in a low(er)-resourced target language. With the new dawn of extremely data hungry pre-trained transformers, the field of cross-lingual knowledge transfer becomes even more effective, since large pre-trained models are not always available for a certain language or task. In this paper, we revisit and demonstrate the strenghts of the anchor extraction approach (initially designed for ELMo) for the task of Bilingual Lexicon Induction (BLI).

The idea of cross-lingual embeddings originally stems from the idea of Mikolov et al. (2013) that vector spaces in different languages share a certain similarity, and a projection can be learned from one language to another. The more recent language models employing contextual embeddings improve upon previous methods of cross-lingual alignment like MUSE (Lample and Conneau, 2019) and VecMap (Artetxe et al., 2018) due to their dynamic nature. Multilingual BERT (mBERT, Devlin et al. (2019)) and XLM (Conneau and Lample, 2019) provide excellent solutions by jointly training for Masked Language Modelling on 104 languages and outperforming previous approaches on zero-shot cross-lingual tasks. However, joint training can be vastly time and computation consuming, and unadaptable for accomodating newer languages after the initial pre-training.

Another interesting approach is proposed by Schuster et al. (2019). They demonstrate that contextual embeddings can be treated as having a static anchor component, and a dynamic context component for every token. This once again enables the static components to be aligned with methods like MUSE. RAMEN (Tran, 2020) proposes a further improvement on the joint training strand of research, by forcing foreign language embeddings to be initialized in the same space as the source language, thus increasing the performance of mBERT and XLM. Artexte et al. (2020) introduce another clever alternative to joint training (mBERT, XLM), by freezing the encoder layers of a transformer after the initial pre-training, and re-learning only the embeddings on a target language. This results in a very similar performance to mBERT while keeping the training time significantly lower.

In this paper, we present a pilot study to investigate a viable approach to minimal supervision cross-lingual transfer of transformer representa-

[1] anonymized

tions from English to three very different target languages, viz. Hindi, Dutch, and Russian. Specifically, we revisit the anchor-based approach of Schuster et al. (2019), which decomposes contextual embeddings into anchors and contexts. To the best of our knowledge, we evaluate for the first time an approach deploying contextual embeddings for the purely lexical task of Bilingual Lexicon Induction, and compare it with two SOTA approaches incorporating static FastText embeddings, being VecMap and MUSE.

2 Investigated Approaches

Alignment of FastText and Word2Vec embeddings using seed dictionaries and Procrustes refinement has been a staple method for cross-lingual adaptation with minimal supervision. However, with transformer architectures being the state of the art for a large majority of downstream tasks after fine-tuning, there was a need for combining two monolingual transformer representations into a joint space. While methods like RAMEN (Tran, 2020) and MonoTrans (Artetxe et al., 2020) accomplish this with additional training of certain parts of the transformers to accomodate multiple languages, in this paper we seek to explore an approach that is robust to multiple architectures, intuitively sound and future-proof for any new large pre-trained language model. The approach in question, referred to as Cross-lingual ELMo (Schuster et al., 2019), theorizes that the average for all contextual embeddings of a word over a large corpus adequately represents a static anchor component of the contextual embeddings, which in theory is sufficient to learn an alignment matrix.

Given a source language s and a target language t, the objective of the classical alignment methods is to learn a transformation,

$$E_{s,t} \approx W^{s \to t} E_{s,s} \tag{1}$$

where $E_{s,s}$ represents the embeddings of the source language in their original space, while $E_{s,t}$ represents the embeddings of the source language, in the target language's multi-dimensional space. Inversely,

$$E_{t,s} \approx W^{t \to s} \tag{2}$$

should also be a possibility. For classical word embeddings like word2vec and FastText, this becomes a simple optimisation problem for an orthogonal matrix W. VecMap achieves this by maximizing for similarity over a sparse seed dictionary (which can be initialized with zero supervision or using identical words if a seed dictionary is not available), and iteratively improving the dictionary and re-learning the alignment after each optimisation step. MUSE achieves the same objective by initializing W using an adversarial objective, where W is optimized such that a discrimnator model is unable to differentiate between the embeddings originating from $E_{t,t}$ and $WE_{s,s}$.

However, the dynamic nature of the embedding spaces E in the case of transformers makes the solutions slightly more complicated and requires some assumptions to simplify the problem. To obtain an approximation of the embedding spaces $E_{s,s}$ and $E_{t,t}$, for a token i in the context c,

$$e_{i,c} = A_i + \hat{e_{i,c}}, \tag{3}$$

where A_i is the fixed Anchor for the token i obtained by averaging embeddings over all available contexts c, while $\hat{e_{i,c}}$ is the additional context component of the embedding. This decomposition means that the complete embedding space $E_{s,s}$ once again can be simplified as a static space as $A_{s,s}$, the space of all anchors for a source language s. A transformation

$$A_{s,t} \approx U^{s \to t} A_{s,s} \tag{4}$$

can therefore be learned with methods like MUSE and VecMap, as for static embeddings.

While this method of alignment for dynamic contextual embeddings was demonstrated to perform very well using ELMo anchors for the task of Dependency Parsing, we further probe the potential of this methodology on the task of Bilingual Lexicon Induction. Even though FastText and Word2Vec-based approaches would perform better on BLI due to the static and purely lexical nature of the task, requiring no contextual complexity, it is interesting to analyze how the computed anchors compare to the simple and elegant FastText embeddings on a purely lexical task. To our knowledge, this is also the first attempt at using BERT-like architectures for the task of BLI, especially in a low-supervision setting.

The contributions of this research can be summarized as follows. We revisit and update the anchor alignment approach from cross-lingual ELMo and test it for the task of BLI for English and three very different target languages with different scripts. We

	EN-HI	HI-EN	EN-NL	NL-EN	EN-RU	RU-EN
FASTTEXT EMBEDDINGS WITH VECMAP						
Full Train Set	0.5679	0.7098	0.8604	0.8467	0.6465	0.8137
1k Supervision	0.4864	0.5268	0.8234	0.766	0.5525	0.7561
FASTTEXT EMBEDDINGS WITH MUSE						
Full Train Set	0.4524	0.5268	0.7834	0.7836	0.6404	0.7765
1k Supervision	0.3348	0.4447	0.7321	0.6968	0.5969	0.7004
ALIGNED ANCHORS WITH VECMAP						
Full Train Set	0.4955	0.5994	0.6382	0.735	0.6210	0.8043
1k Supervision	0.3620	0.2997	0.230	0.386	0.3276	0.5940

Table 1: BLI Results for the six language pairs, with English (EN), Hindi (HI), Dutch (NL) and Russian (RU) as source and target language.

demonstrate the flexibility of the approach by using multiple architectures from the BERT family in our experiments, analyze the post-alignment anchors and discuss recurring issues across all languages.

3 Experimental Setup

We perform Bilingual Lexicon Induction experiments for three languages Hindi (HI), Dutch (NL), and Russian (RU). Using English (EN) as a source language and a target language for each language. All datasets have been derived from the MUSE bilingual dictionaries[2]. Because our intention is to evaluate contextual models, the respective MUSE train and test sets had to be reduced to accommodate for the smaller sub-word based vocabularies compared to classical FastText or word2vec variants. Using the full dictionaries would be misleading, since, for example, for Russian, our model was only able to use around 3500 samples for training, as compared to the 5000 available in the full train set. To keep the comparisons consistent, we evaluated the two methods incorporating static FastText embeddings (VecMap and MUSE) on the reduced train/test sets as well, and make the reduced dictionaries available[3] for reproducibility.

For our proposed approach, anchors have been derived from a variety of different transformer architectures to demonstrate the flexibility of the anchor extraction methodology. While for English and Hindi, anchors extracted from more standard BERT-uncased models were used, we relied on RuBERT (Kuratov and Arkhipov, 2019) for Russian, which is a cased BERT model, and on Robbert (Delobelle et al., 2020) for Dutch, which is a RoBERTa-based architecture.

For all three languages, we use English both as a source and target language. Two sets of experiments have been performed for each language pair: one with the completely available training set, and a second one where only 1000 samples are available for supervision. We use FastText vectors aligned with the same hyperparameters as the anchors, using VecMap and MUSE for comparisons.

To evaluate, we calculate accuracy, which measures whether the predicted translation in a certain target language was also among the gold standard translations in that language. As our primary objective is to evaluate whether an anchor-based approach is feasible for BLI relying on transformer-based architectures, an error analysis is also performed.

4 Results

The accuracy scores for all experiments are shown in Table 1. The anchor alignment methods outperform the MUSE alignments for some setings for Hindi and Russian, but fails to compete with the SOTA VecMap alignments using FastText. A reason why FastText embeddings align significantly better can be attributed to the isomorphism assumptions. Vulić et al. (2020) pointed out that two sets of embeddings are more likely to be isomorphic given the same amount of training data, time and parameters. This makes FastText very robust since embeddings for all the languages are trained in an identical fashion. Intuitively this isomorphism assumption should also cause a major issue for Dutch where we use a different architecture (RoBERTa) for alignment, and for Russian where a cased model was used to align with an uncased model. Nevertheless, the performances indicate that the anchor extraction is robust to these issues.

[2] https://paperswithcode.com/dataset/muse

[3] anonymised

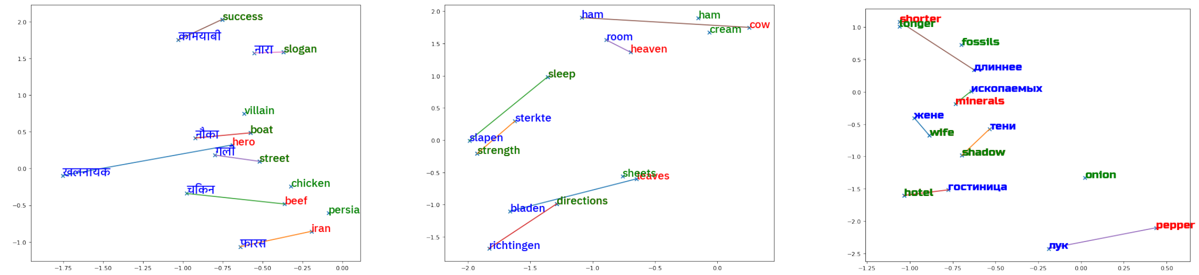

Figure 1: Illustration of Hindi, Dutch and Russian example words (blue), respectively, that are correctly (green) and incorrectly (red) aligned according to the gold standard.

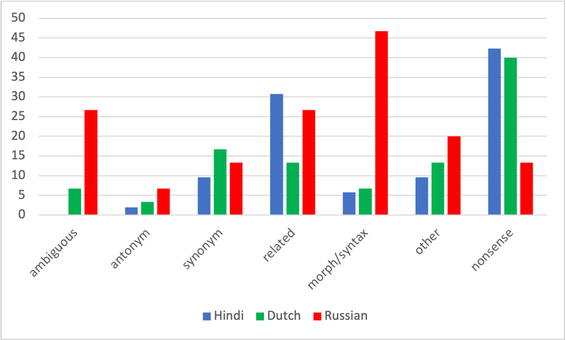

Figure 2: Distribution of error types per language (percentagewise)

We can make a few key observations based on the results. Firstly, we note that with the anchor-based approach, the transfer to English is significantly harder than relying on English as the source language. Another noteworthy outcome is that the drop-off of performance for the 1,000 training samples experiments seems to be consistently higher for the anchor alignments compared to FastText. This again could be attributed to the larger vocabulary of FastText allowing the alignment refinement steps to have a better understanding of the embedding space.

In order to get more insights into the ouput of our approach, we performed a qualitative error analysis on the output of the first 100 instances of the test sets in all three languages (with English as the source). Interestingly, we found that even though these three language are far apart, they exhibit similar errors. Figure 2 represents, percentagewise, which errors were made as judged by native experts of the respective languages. As can be observed, the largest error category in Hindi and Dutch constitutes nonsensical words, a typical problem of BERT-based architectures, whereas for Russian especially morphologically/syntax-related errors prevail (the latter has mostly to do with different cases or inflections of nouns, a typical difficulty of the Russian language). Looking at the other error types, we observe that these have mostly to do with semantics (antonyms, synonyms, polysemous words), of which words that are somewhat semantically related (example EN-HI: 'chicken' was translated as *elephant*, example EN-RU: 'promise' was translated to *hope*, example EN-DU: 'inches' was translated by *meters*, which is actually the Dutch standard distance metric) seem to pose a larger problem, especially in Hindi and Russian and to a lesser extent in Dutch.

In Figure 1 we, also attempt to visualize some selected embeddings correctly aligned, and some incorrectly aligned using PCA, for all three language pairs. The embeddings in green represent the correct translations in the target language, while the ones in red are incorrectly aligned. The visualizations demonstrate (again) that a lot of the mistakes can be attributed to semantics, as well as ambiguity in the test set (eg. 'bladen' in Dutch can be interpreted as both 'sheets' (*of paper*) and 'leaves' (*of tree*), but only 'sheets' is accepted by the gold standard test set).

In conclusion, we demonstrated that extracting static anchors from transformers is a viable method for low supervision Bilingual Lexicon Induction. The aligned anchors were able to outperform one of the SOTA approaches (MUSE) for some settings, and show great promise in terms of flexibility and adaptability for different languages and architectures in the BERT family. An in-depth error analysis and visualization of the aligned anchors shows that a lot of the mistakes can be attributed to semantic or syntatic (primarily Russian) misunderstandings. In future research, we will investigate the impact of using a larger corpus for extracting the anchors, and evaluate the approach for other more complex tasks like NLI and Question Answering.

References

M. Artetxe, G. Labaka, and E. Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

A. Conneau and G. Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

G. Lample and A. Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop Papers*.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Ke M. Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *CoRR*, abs/2002.07306.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic?