Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study

Saurabh Kulshreshtha* Department of Computer Science University of Massachusetts Lowell skul@cs.uml.edu

José Luis Redondo-García Amazon Alexa Cambridge, UK jluisred@amazon.com

Ching-Yun Chang Amazon Alexa Cambridge, UK cychang@amazon.com

Abstract

Multilingual BERT (mBERT) has shown reasonable capability for zero-shot cross-lingual transfer when fine-tuned on downstream tasks. Since mBERT is not pre-trained with explicit cross-lingual supervision, transfer performance can further be improved by aligning mBERT with cross-lingual signal. Prior work proposes several approaches to align contextualised embeddings. In this paper we analyse how different forms of cross-lingual supervision and various alignment methods influence the transfer capability of mBERT in zero-shot setting. Specifically, we compare parallel corpora vs. dictionary-based supervision and rotational vs. fine-tuning based alignment methods. We evaluate the performance of different alignment methodologies across eight languages on two tasks: Name Entity Recognition and Semantic Slot Filling. In addition, we propose a novel normalisation method which consistently improves the performance of rotation-based alignment including a notable 3% F1 improvement for distant and typologically dissimilar languages. Importantly we identify the biases of the alignment methods to the type of task and proximity to the transfer language. We also find that supervision from parallel corpus is generally superior to dictionary alignments.

1 Introduction

Multilingual BERT (mBERT) (Devlin et al., 2019) is the BERT architecture trained on data from 104 languages where all languages are embedded in the same vector space. Due to the multilingual and contextual representation properties of mBERT, it has gained popularity in various multilingual and cross-lingual tasks (Karthikeyan et al., 2020; Wu and Dredze, 2019). In particular, it has demonstrated good zero-shot cross-lingual transfer performance on many downstream tasks, such as Document Classification, NLI, NER, POS tagging, and Dependency Parsing (Wu and Dredze, 2019), when the source and the target languages are similar.

Many experiments (Ahmad et al., 2019) suggest that to achieve reasonable performance in the zero-shot setup, the source and the target languages need to share similar grammatical structure or lie in the same language family. In addition, since mBERT is not trained with explicit language signal, mBERT's multilingual representations are less effective for languages with little lexical overlap (Patra et al., 2019). One branch of work is therefore dedicated to improve the multilingual properties of mBERT by aligning the embeddings of different languages with cross-lingual supervision.

Broadly, two methods have been proposed in prior work to induce cross-lingual signals in contextual embeddings: 1) Rotation Alignment as described in Section 2 aims at learning a linear rotation transformation to project source language embeddings into their respective locations in the target language space (Schuster et al., 2019b; Wang et al., 2019; Aldarmaki and Diab, 2019); 2) Fine-tuning Alignment as explained in Section 3 internally aligns language sub-spaces in mBERT through tuning its weights such that distances between embeddings of word translations decrease while not losing the informativity of the embeddings (Cao et al., 2020). Additionally, two sources of cross-lingual signal have been considered in literature to align languages: parallel corpora and bilingual dictionaries. While the choice of each alignment method and source of supervision have a variety of advantages and disadvantages, it is unclear how these affect the performance of the aligned spaces across languages and various tasks.

In this paper, we empirically investigate the effect of these cross-lingual alignment methodologies and applicable sources of cross-lingual super-

* Work done during an internship at Amazon.

vision by evaluating their performance on zero-shot Named Entity Recognition (NER), a structured prediction task, and Semantic Slot-filling (SF), a semantic labelling task, across eight language pairs.

The motivation for choice of these tasks to evaluate are two-fold: 1. Prior work has already studied alignment methods on sentence level tasks. Cao et al. (2020) show the effectiveness of mBERT alignment methods on XNLI (2018). 2. Word-level tasks do not benefit from more pre-training unlike other language tasks that improve by simply supplementing with more pre-training data. In experiments over the XTREME benchmark, Hu et al. (2020) find that transfer performance improves across all tasks when multilingual language models are pre-trained with more data, with the sole exception of word-level tasks. They note that this indicates current deep pre-trained models do not fully exploit the pre-training data to transfer to word-level tasks. We believe that NER and Slot-filling tasks are strong candidate tasks to assess alignment methods due to limited cross-lingual transfer capacity of current models to these tasks.

To the authors' knowledge, this is the first paper exploring the comparison of alignment methods for contextual embedding spaces: rotation vs. fine-tuning alignment and two sources of cross-lingual supervision: dictionary vs. parallel corpus supervision on a set of tasks of structural and semantic nature over a wide range of languages. From the results, we find that parallel corpora are better suited for aligning contextual embeddings. In addition, we find that rotation alignment is more robust for primarily structural NER downstream tasks while the fine-tuning alignment is found to improve performance across semantic SF tasks. In addition, we propose a novel normalisation procedure which consistently improves rotation alignment, motivated by the structure of mBERT space and how languages are distributed across it. We also find the effect of language proximity on transfer improvement for these alignment methods.

2 Rotation-based Alignment

Mikolov et al. (2013) proposed to learn a linear transformation $W_{s \to t}$ which would project an embedding in the source language e_s to its translation in the target language space e_t, by minimising the distances between the projected source embeddings and their corresponding target embeddings:

$$\min_{W \in R^{d \times d}} \|WX_s - X_t\| \tag{1}$$

X_s and X_t are matrices of size $d \times K$ where d is the dimensionality of embeddings and K is the number of parallel words from word-aligned corpora, or word pairs from a bilingual dictionary between the source and target languages. Further work Xing et al. (2015) demonstrated that restricting W to a purely rotational transform improves cross-lingual transfer across similar languages. The orthogonality assumption reduces Eq.(1) into the so-called Procrustes problem with the closed form solution:

$$W = UV^T, \tag{2}$$

$$\text{where } U\Sigma V^T = SVD\left(X_t X_s^T\right) \tag{3}$$

and the SVD operator stands for Singular Value Decomposition.

2.1 Language Centering Normalization

A purely rotational transformation can align two embedding spaces only if the two spaces are roughly isometric and are distributed about the same mean. In case the two embedding distributions are not centered around the same mean, meaning the two spaces have little overlap and are shifted by a translation offset in the space, they cannot be aligned solely through rotation.

Since the linear transformation $W_{s \to t}$ derived from solving the Procrustus problem only rotates the vector space, it assumes the embeddings of two languages are zero-centered. However Libovický et al. (2019) observe that languages distributions in mBERT have distinct and separable centroids and different language families have well separated sub-spaces in the mBERT embedding vector space. To address this discrepancy, we propose a new normalisation mechanism which entails:

Step 1. Normalising the embeddings of both languages so that they have zero mean:

$$\hat{X}_s = X_s - \bar{X}_s \text{ and } \hat{X}_t = X_t - \bar{X}_t \tag{4}$$

where $\bar{X}_s$ and $\bar{X}_t$ are centroids of source and target embeddings X_s and X_t; and $\hat{X}_s$ and $\hat{X}_t$ are mean-centered source and target language embeddings their rows correspond to word translations. Next, $\hat{X}_s$ and $\hat{X}_t$ are used to compute the transformation matrix $\hat{W}_{s \to t}$ by solving Eq.(2) and Eq.(3).

Step 2. During training a downstream task, embedding of a source language word e_s needs to be re-centered, rotated and finally translated to the target language subspace to derive the projection e_{t^*}:

$$e_{t^*} = \hat{W}_{s \to t}(e_s - \bar{X}_s) + \bar{X}_t \tag{5}$$

This helps the task specific model, particularly in zero-shot setting, by projecting the source language task data to the same locality as the target language.

2.2 Supervision Signals for Rotation Alignment

In this section we describe how existing work utilises two different cross-lingual signals, bilingual dictionaries and parallel corpora, to supervise rotation alignment. Additionally, we analyse the advantages and disadvantages of the two choices.

2.2.1 Bilingual Dictionary Supervision

In order to utilise a bilingual dictionary to supervise the embedding alignment, each word in the dictionary needs to have a single representation. However the same word can have many representations in the contextualised language model vector space depending on the context it occurs in. Schuster et al. (2019b) observes that the contextual embeddings of the same word form a tight cluster - word cloud, the centroid of this word cloud is distinct and separable for individual words. They further propose that centroid of a word cloud can be considered as the context-independent representation of a word, called average word anchor. These word anchors are computed by averaging embeddings over all occurrences of a word in a monolingual corpora, where words occur in a variety of contexts. Formally the mBERT embedding of a source language word s_m in context c_h is denoted as e_{s_m,c_h}. If this word occurs a total of p times in the monolingual corpus, that is in contexts $c_1, c_2, ...c_p$, the anchor word embedding A_{s_m} for word s_m across all the contexts is the average:

$$A_{s_m} = \frac{\sum_{h=1}^{p} e_{s_m,c_h}}{p} \quad (6)$$

Average word anchor pair $(A^i_{s_m}, A^i_{t_{m^*}})$, where i is the mBERT layer, for all word pairs from the dictionary (s_m, t_{m^*}) form the rows of matrices X^i_s and X^i_t respectively, which are then used to solve Eq.(2) and Eq.(3), resulting in an alignment transformation matrix $W^i_{s \to t}$.

However, there are limitations to this approach. Zhang et al. (2019) found that the word cloud of multi-sense words, such as the word "bank", which can mean either the financial institution or the edge of a river depending on the context, are further composed of clearly separable clusters, for every word sense. Averaging over multiple contextual embeddings infers losing certain degree of contextual information at both the source and target language words. Figure 1a visualises word anchor calculation and also highlights this limitation. On the other hand, one of the advantages of this method is that bilingual dictionaries are available for even very low resource languages.

2.2.2 Parallel Corpus Supervision

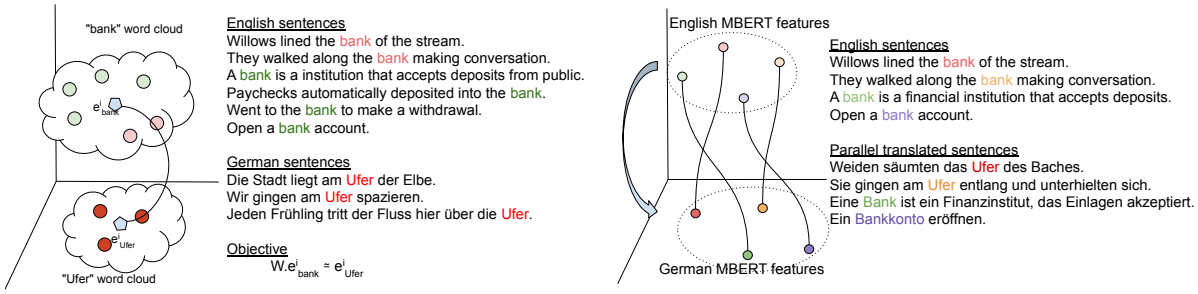
Word-aligned parallel sentences can be utilised as a source of cross-lingual signal to align contextual embeddings (Aldarmaki and Diab, 2019; Wang et al., 2019). Given a parallel corpora, s_m and t_{m^*} are aligned source and the target language words appearing in context c_h and c_{h^*}, respectively. The parallel word embedding matrices X^i_s and X^i_t for mBERT layer-i are composed from the contextual embeddings $e^i_{s_m,c_h}$ and $e^i_{t_{m^*},c_{h^*}}$ respectively, and are used to solve Eq.(2) and Eq.(3) to derive an alignment transformation matrix $W^i_{s \to t}$.

Figure 1a and 1b illustrate how parallel supervision is more suited to align contextual embeddings compared to dictionary supervision where multiple senses of a word are compressed into a single word anchor. However, parallel corpora rarely come with word-alignment annotations that are often automatically generated by off-the-shelf tools such as *fast_align* (Dyer et al., 2013), which can be noisy. It is worth noting that word alignment error rate of an off-the-shelf tool drops when number of parallel sentences increases, therefore parallel corpus supervision is favourable for languages where more parallel data is available.

3 Fine-tuning Alignment with Parallel Corpora

Rotation alignment has a strong assumption that the two language spaces (or sub-spaces in case of mBERT) are approximately isometric (Søgaard et al., 2018). Patra et al. (2019) reported that the geometry of language embeddings becomes dissimilar for distant languages, and the isometry assumption degrades the alignment performance in such cases. In addition, as explained in Section 2.1 rotation alignment alone cannot achieve effective mapping when two languages spaces have separate centroids. Therefore, next we consider existing work to non-linearly align two language spaces.

Cao et al. (2020) proposed to directly align languages within mBERT model through fine-tuning.



(a) Alignment with dictionary.

(b) Alignment with parallel sentences.

Figure 1: In Figure 1a, contextual embeddings of the word "bank" get averaged across all word senses noted by different colors into single word anchor embedding. Figure 1b illustrates supervision from parallel corpora where word-alignments correspond to translation in similar context noted by similar colors (lighter for English), this provides more fine-grained supervision for contextualised alignment of mBERT.

The objective of the fine-tuning is to minimise the distance between the two contextual representations of an aligned word pair in parallel corpora:

$$L^i_{align} = min \sum_{m,m^*} \|e^i_{s_m} - e^i_{t_{m^*}}\| \quad (7)$$

However, fine-tuning with only the above objective would led to lose the semantic information in mBERT learnt during pre-training, since a trivial solution to the Eq.(7) can be simply to make all the embeddings equal. To deal with this, Cao et al. (2020) also proposed a regularisation loss that does not allow the embedding of a source language word to stray too far away from its original location $\mathbf{e}^i_{s_m}$ in the pre-trained mBERT model, namely:

$$L^i_{regularise} = min \sum_{m} \|e^i_{s_m} - \mathbf{e}^i_{s_m}\| \quad (8)$$

Note that $\mathbf{e}^i_{s_m}$ is generated from a copy of the original pre-trained mBERT model where the parameters are kept frozen. Both of the alignment and the regularization losses are combined and jointly optimised in order to align the two language subspaces while maintaining informativity of embeddings:

$$L_{finetune} = min \sum_{i=n_s}^{n_e} L^i_{align} + L^i_{regularise} \quad (9)$$

Here n_s to n_e is the range of mBERT layers aligned. We experimented with two variants of the fine-tuning approach: 1) moving target language towards source language while keeping the source embeddings approximately fixed through the regularization term in Eq.(8); 2) moving the source language embeddings towards the target space while keeping the target language space relatively fixed, then the regularisation loss changes to:

$$L^i_{regularise} = min \sum_{m^*} \|e^i_{t_{m^*}} - \mathbf{e}^i_{t_{m^*}}\| \quad (10)$$

4 Experimental Setup

In this section, we firstly describe the resources and implementation details of the alignment methods followed by the zero-shot NER and SF tasks used to evaluate the alignments. In addition, we briefly explain the datasets used in the experiments.

4.1 Learning Alignments

Our baseline model is a pre-trained mBERT* – 12 transformer layers, 12 attention heads, 768 hidden dimensions – denoted as *mBERT Baseline*. When a word is tokenised into multiple subwords by the tokeniser, we average their corresponding subword embeddings to obtain embedding for the word. Following Wang et al. (2019) we collect 30k parallel sentences for each of the language pairs from publicly available parallel corpora. For the European languages, German, Italian, Spanish and Dutch, the Europarl corpus (Koehn, 2005) is used; for Hindi, Turkish and Thai, the OpenSubtitles corpus (Lison and Tiedemann, 2016) is used; for Armenian the parallel sentences are extracted from the QED Corpus (Abdelali et al., 2014). We obtain contextual and average anchor embeddings described in Section 2.2.1 by passing the corpora described above through pre-trained mBERT.

We use the bilingual dictionaries provided with the MUSE framework (Lample et al., 2018) as the source for dictionary supervision. As for the parallel corpus supervision, since none of the collected parallel sentences contains word-level alignment information, we utilise *fast_align* (Dyer et al., 2013) to automatically derive word alignment signals.

*Available for download at: https://github.com/google-research/bert/blob/master/multilingual.md

For the rotation alignment, we compute four independent transformation matrices for each of the last four transformer layers similar to Wang et al. (2019). We use *RotateAlign* and *NormRotateAlign* to refer the rotation alignment learnt without and with the proposed language centering normalisation, respectively. To be consistent, for the fine-tuning alignment we align the word representations in the last four transformer layers of the mBERT model, denoted as *FineTuneAlign*.

4.2 Evaluation of the Alignments

We evaluate the learnt alignments using two downstream tasks: Named Entity Recognition (NER) and Semantic Slot Filling (SF), both of which aim to predict a label for each token in a sentence. NER is a more structural task with fewer entity types and involves less semantic understanding of the context compared to SF. Examples of the tasks can be found in Table 2.

We use the same model architecture and hyper-parameters as Wang et al. (2019), two BiLSTM layers followed by a CRF layer, where learning rate is set to 10^{-4} for European languages and 10^{-5} for the other languages determined by the validation set. In order to measure the effectiveness of a learnt alignment, all the experiments are conducted with zero-shot settings similar to Wang et al. (2019), where the source language data is first transformed to the target language space and then used to train a BiLSTM-CRF model. The target language validation set is used for hyper-parameter tuning and reporting the evaluation results. For each experiment we report F1 scores averaged over 5 runs.

4.3 NER and SF Datasets

We use the following four families of datasets, each of which has the same set of labels. A summary of the datasets can be found in Table 1. Example utterances and annotations and shown in Table 2.

CoNLL-NER: This includes CoNLL 2002, 2003 NER benchmark task (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) containing entity annotations for news articles in English, German, Spanish and Dutch. We also include in this family PioNER[†] (Ghukasyan et al., 2018), a manually annotated dataset in Armenian, which is typographically different from the other languages in this family. In this dataset-family, target language data is sourced from local news articles, and not generated through translation from source data.

ATIS-SF: ATIS Corpus (Price, 1990) is an English dataset containing conversational queries about flight booking. Upadhyay et al. (2018) manually translated a subset of the data into two languages, Turkish and Hindi, along with crowd-sourced phrase-level annotations.

FB-SF: Schuster et al. (2019a) introduced Multilingual Task-Oriented Dialog Corpus in English, Spanish and Thai across three domains: weather, alarm and reminders, where Spanish and Thai data were manually translated and annotated from a subset of the English data.

SNIPS-SF: A multi-domain slot-filling dataset in English released by Coucke et al. (2018). Bellomaria et al. (2019) automatically translated this dataset into Italian, and then manually labelled the translation where entities were substituted by Italian entities collected from the Web.

5 Results and Analysis

The evaluation results of each alignment method on the downstream NER and SF tasks are reported in Table 3 and Figure 2. In addition to the *mBERT Baseline* and for comparison purposes, we also list relevant results found in literature (Wu and Dredze, 2019; Wang et al., 2019; Upadhyay et al., 2018; Schuster et al., 2019a; Bellomaria et al., 2019) that have been evaluated on the same datasets.

5.1 mBERT Baseline and Language Proximity

mBERT Baseline numbers can be indicative of how well languages are already aligned in the mBERT space. High zero-shot scores for German, Dutch, Spanish and Italian indicate that European languages are extremely well aligned to English in mBERT. However, distant languages such as Thai and Turkish, which belong to different language families (Kra–Dai and Turkic) than English, have poor alignment with low F1 scores of 9.58 and 21.15, respectively. Finally, moderately distant languages such as Armenian and Hindi, which fall within the larger Indo-European language family, have moderate alignment with English with scores of 62.38 and 50.84, respectively.

[†]PioNER data only has PER, LOC and ORG labels and does not contain MISC.

Datasets	Task	Translated	Language & Train/Dev/Test Size	# Slot Types	Domains
CoNLL(2002; 2003) PioNER[1](2018)	NER	No	**en** 14,987 / 3,466 / 3,684 de 12,705 / 3,068 / 3,160 es 8,323 / 1,915 / 1,517 nl 15,806 / 2,895 / 5,195 hy 5,964 / 1,491 / 2,529	4	News Articles
ATIS(1990) ATIS-HI,TK(2018)	SF	Yes	**en** 4,478 / 500 / 893 hi 600 / 893 / 893 tk 600 / 715 / 715	63	Air Travel
FB(2019a)	SF	Yes	**en** 30,521 / 4,181 / 8,621 es 3,617 / 1,983 / 3,043 th 2,156 / 1,235 / 1,692	11	Weather, Alarm, Reminder
SNIPS(2018) Almaware-SLU(2019)	SF	Yes	**en** 13,084 / 700 / 700 it 1,400 / 700 / 700	39	Music, Restaurants, TV, Movies, Books, Weather

Table 1: Summary of NER and SF dataset families. English marked in bold is treated as the source language.

CoNLL-NER	[U.N.]$_{ORG}$ official [Ekeus]$_{PER}$ heads for [Baghdad]$_{LOC}$.
ATIS-SF	show the [latest]$_{flight\_mod}$ flight from [denver]$_{fromloc.city\_name}$ to [boston]$_{toloc.city\_name}$
FB-SF	do you have [wednesday's]$_{datetime}$ [weather forecast]$_{weather\_noun}$ for [half moon bay]$_{location}$
SNIPS-SF	add this [track]$_{music\_item}$ to [my]$_{playlist\_owner}$ [global funk]$_{playlist}$

Table 2: Examples from the datasets.

5.2 mBERT Baseline vs./ Rotation Alignment

RotateAlign improves performance by 19% absolute for ATIS-Turkish, going from baseline of 21.15 to 38.18 in F1 score. For ATIS-Hindi the performance improves from 50.84 to 57.86 F1 (7 points), and 4% absolute for the PioNER-Armenian from 62.38 to 66.56. These numbers show how *RotateAlign* can improve performance over *mBERT Baseline* for moderately-close languages such as Hindi, Turkish and Armenian, while there is only around 1 point improvement for European languages. This implies that Hindi, Turkish and Armenian subspaces are geometrically similar to English, however they are misaligned in terms of rotation in *mBERT Baseline*.

However, in the case of Thai, which is a distant language from English, *RotateAlign* does not improve performance over the *mBERT Baseline*. This suggests that Thai and English's embedding spaces are structurally dissimilar.

5.3 Rotation Alignment with vs./ without Language Centering Normalisation

Applying the proposed language centering normalisation in Section 2.1 before performing the rotation alignment, namely *NormRotateAlign* in Table 3, is found to further improve downstream performance across all tasks and languages. The improvement over *RotateAlign* is up to 3% absolute F1 for Thai, around 1% absolute for moderately closer languages like Hindi, Turkish and Armenian, and around 0.5% absolute F1 for closer target languages such as German. Note that Thai, which does not benefit from rotation alignment alone, improves by an average of 2.3 points after applying the normalisation. These results corroborate that language families that are further away from each other have more separable sub-spaces in the *mBERT Baseline*, and bringing the language distributions closer helps the downstream task's performance.

5.4 Parallel Corpus vs./ Dictionary Supervision

Amongst the cases where *RotateAlign* improves performance over the *mBERT Baseline*, parallel-corpus supervised *RotateAlign* is superior to dictionary supervision, with the exception of Hindi. This could be explained by the fact that word anchors are independent of multiple word senses, thereby the cross-lingual signal is poorer compared to parallel word alignments. This is in line with observations from Zhang et al. (2019).

5.5 Rotation vs./ Fine-tuning Alignment

From Table 3 and Figure 2 we can see that *FineTuneAlign* explained in Section 3 improves performance over *RotateAlign* for semantic tasks (SF), with the only exception of ATIS-Hindi. On the other hand, *FineTuneAlign* underperforms *RotateAlign* for structural tasks (NER), and in some cases even fall behind *mBERT Baseline*. Note that we notice no clear trend between *FineTuneAlign*$_{src \to tgt}$ and *FineTuneAlign*$_{tgt \to src}$.

Dataset-Task	CoNLL-NER				ATIS-SF		FB-SF		SNIPS-SF
Transfer Pair	en to de	en to nl	en to es	en to hy	en to hi	en to tk	en to es	en to th	en to it
Baselines from Literature									
mBERT (Wu and Dredze, 2019)	69.56	77.75	74.96	-	-	-	-	-	-
mBERT Rotation Alignment: Parallel (Wang et al., 2019)	70.54	79.03	75.77	-	-	-	-	-	-
BERT, 1400 Target Language Train (Bellomaria et al., 2019)†	-	-	-	-	-	-	-	-	83.04
Non-contextual Zero-shot Baseline (Upadhyay et al., 2018)*	-	-	-	-	∼40	∼40	-	-	-
Translate train (Schuster et al., 2019a)‡	-	-	-	-	-	-	72.87	55.43	-
Our Experiments									
mBERT Baseline	66.15	77.55	74.80	62.38	50.84	21.15	74.66	9.58	76.70
RotateAlign$_{dict}$	67.20	78.07	75.08	-	57.32	31.46	73.28	9.23	76.51
NormRotateAlign$_{dict}$	68.56	78.53	75.22	-	**57.86**	33.62	74.52	12.38	76.82
RotateAlign$_{parallel}$	70.48	79.52	75.84	65.31	52.24	37.38	73.57	9.12	77.70
NormRotateAlign$_{parallel}$	**71.23**	**79.90**	**75.93**	**66.56**	53.03	38.18	74.73	11.88	77.87
FineTuneAlign$_{tgt\rightarrow src}$	70.25	77.10	73.92	63.53	51.35	**45.98**	73.44	13.45	77.74
FineTuneAlign$_{src\rightarrow tgt}$	66.91	77.21	74.49	62.29	50.51	39.43	**80.90**	**20.77**	**80.21**

Table 3: Performance (F1 score) of the alignment methods on the zero-shot NER and SF tasks. Top scores within our experiments are marked in bold. No results are reported for Armenian dictionary alignments since English-Armenian dictionary was available in the MUSE framework. † Bellomaria et al. (2019) use 1400 Italian instances as part of the training data. * Numbers read from a chart in the paper. ‡ Schuster et al. (2019a) uses a machine translation model to translate this dataset and word alignments generated by attention weights to infer annotation.

FineTuneAlign$_{src\rightarrow tgt}$ improves over the best rotation alignment *NormRotateAlign*$_{parallel}$ by 7.8% absolute for the ATIS-Turkish task from 38.18 to 45.98. It significantly outperforms *mBERT Baseline* by 24 points. For FB-Thai *FineTuneAlign*$_{src\rightarrow tgt}$ surpasses *NormRotateAlign*$_{dict}$ by 8.39% absolute F1 from 12.38 to 20.77, 11 points higher than *mBERT Baseline*. For FB-Spanish we observe an improvement from 74.73 to 80.90 (6% absolute) compared to *RotateAlign* and similarly +6 points compared to *mBERT Baseline*. For SNIPS-Italian, *FineTuneAlign* improves performance over *NormRotateAlign* from 77.87 to 80.21 (2.5 points) and is 3.5 points better than *mBERT Baseline*.

All SF tasks considered are generated by translation from the source language data. This may indicate that the fine-tuning approach performs better than rotation-based methods for translated datasets, where there is high correlation between utterance structure of training data in source language and evaluation data in target language. On the other hand, rotation-based alignments generalise better when the downstream target sentence distribution is dissimilar from the source sentence distribution, as is the case for non-translated NER tasks.

5.6 Aligned Source Language vs./ Target Language Training

FineTuneAlign$_{src\rightarrow tgt}$ achieves top F1 score of 80.21 on SNIPS-Italian dataset which is not far from the score of 83 from a BERT-based model trained on 1400 manually-annotated Italian utterances (2019). Also, our best alignment score of 80.90 for FB-Spanish (*FineTuneAlign*$_{src\rightarrow tgt}$) surpasses translate-train baseline (2019a) where the annotations are automatically inferred from a NMT model. This suggests that for closer target languages, fine-tuning based alignment are not far behind from unaligned models trained on additional target language labelled examples.

Performance improvement from fine-tuning alignment for translated datasets should not be attributed to superficial transfer of entity information from source language. An evidence to support this claim is the strong performance on the SNIPS Italian-SF dataset, which has been translated from SNIPS dataset (Bellomaria et al., 2019), where English entities have been replaced with Italian entities collected from the Web during dataset preparation. Therefore, during validation, the model came across utterances with similar structure but different entities, which shows that improvement from fine-tuning alignment is largely independent of language specific entity memorisation.

6 Related Work

Aldarmaki and Diab (2019) propose to align ELMo embeddings (Peters et al., 2018) with word-level and sentence-level alignments. They compare the aligned ELMo with static character-level embeddings with similar alignments.

Cao et al. (2020) originally proposed fine-tuning alignment of mBERT language sub-spaces. They claim these methods are strictly stronger to rotation alignments methods based solely on zero-shot experimentation on XNLI task (Conneau et al., 2018), a semantic sentence-level classification task gener-

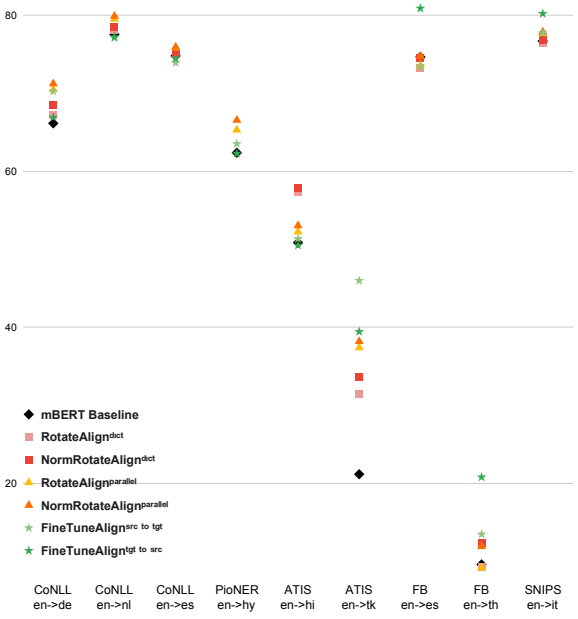


Figure 2: Trend of improvement from various alignment methods. Rotation alignment improves performance for NER, while fine-tuning alignment is found to be better for SF tasks. Improvements increase initially with distance between source and target languages and diminish for distant languages.

ated through translation from source language. On the contrary, we observe that fine-tuning does not improve performance across all tasks, particularly structural tasks, where utterance structure changes and there is higher incidence of domain shift. This raises the question whether translated datasets are biased to fine-tuning alignment, and whether such datasets are a good evaluation test-bed for general cross-lingual transfer.

Wang et al. (2019) applies rotational alignment to mBERT and reports results on CoNLL NER tasks, however the main focus of their work is on the overlap of static bilingual embeddings. They do not extend similar analysis on contextualised embeddings. In our work, drawing from the observations made by Libovický et al. (2019) on the distribution of languages in mBERT space, we propose a normalization mechanism to increase the overlap of two languages distributions prior to computing rotational alignment.

Schuster et al. (2019b) originally proposed dictionary supervision to align ELMo with rotational transform. They claim supervision from dictionary is superior to using parallel word aligned corpora, however they do not substantiate these through comparative experiments. We observe that parallel corpus supervision is stronger than dictionary supervision possibly because of considering contextual alignment.

7 Conclusion

In this paper, we investigate cross-lingual alignment methods for multilingual BERT. We empirically evaluate their effect on zero-shot transfer for downstream tasks of two types: structural NER and semantic Slot-filling, across a set of diverse languages. Specifically, we compare rotation alignment and fine-tuning cross-lingual alignment. We compare the effect of dictionary and parallel corpora supervision across all tasks. We also propose a novel normalisation technique that improves state-of-the-art performance on zero-shot NER and Semantic Slot-filling downstream tasks, motivated by how languages are distributed across the mBERT space. Our experimental settings cover four datasets families (one for NER and three for SF) across eight language pairs.

Key findings of this paper are as follows: (1) rotation-based alignments show large performance improvements (up to +19% absolute for Turkish ATIS-SF) on moderately close languages, only a small improvement for very close target languages and no improvement for very distant languages; (2) we propose a novel normalisation which centers language distributions prior to learning rotation maps and is consistently shown to improve rotation alignment across all tasks particularly for Thai, by up to 3% absolute; (3) rotational alignments are more robust and generalise well for structural tasks such as NER which may have higher utterance variability and domain shift; (4) supervision from parallel corpus generally leads to better alignment than dictionary-based, since it offers the possibility of generating contextualised alignments; (5) fine-tuning alignment improves performance for semantic tasks such as slot-filling where the source language data has minimal shift in utterance structure or domain from target language data and particularly improves performance for extremely distant languages (up to +8.39% absolute higher for Thai FB-SF) compared to rotation alignment; (6) for close languages and tasks with similar utterance structure, zero-shot fine-tuning alignment is competitive versus unaligned models trained on additional annotated data in target language.

This work aims to pave the way for optimising language transfer capability in contextual multilingual models. In the future, we would like to further

investigate patterns in the embedding space and apply alignment methods into specific regions of the multilingual hyperspace to obtain more tailor-suited alignments between language pairs. We would also like to evaluate zero-shot capabilities of alignments when applied to other language tasks.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

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 *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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Hanan Aldarmaki and Mona Diab. 2019. Context-Aware Cross-Lingual Mapping. 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 3906–3911, Minneapolis, Minnesota. Association for Computational Linguistics.

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almawave-SLU: A new dataset for SLU in Italian.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Tsolak Ghukasyan, Garnik Davtyan, Karen Avetisyan, and Ivan Andrianov. 2018. pioNER: Datasets and Baselines for Armenian Named Entity Recognition. *2018 Ivannikov Ispras Open Conference (ISPRAS)*, pages 56–61.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *CoRR*, abs/2003.11080.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. *ArXiv*, abs/1912.07840.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How Language-Neutral is Multilingual BERT?

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2015: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *International Conference on Language Resources and Evaluation*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4168.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

P. J. Price. 1990. Evaluation of Spoken Language Systems: the ATIS Domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019a. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. 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 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019b. 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.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

S. Upadhyay, M. Faruqui, G. Tür, H. Dilek, and L. Heck. 2018. (Almost) Zero-Shot Cross-Lingual Spoken Language Understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2019. Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Zheng Zhang, Ruiqing Yin, Jun Zhu, and Pierre Zweigenbaum. 2019. Cross-Lingual Contextual Word Embeddings Mapping With Multi-Sense Words In Mind.