Cross Lingual Contextual Embedding Alignment

Nick Schoelkopf
September 23, 2021

Outline

1. Introduction
    a. Embeddings Overview
    b. Static Embedding Alignment
    c. Multilingual BERT + Contextualized Embeddings
2. Papers
3. Conclusion
4. Discussion

Papers Covered

Paper 1: Word Alignment by Fine-tuning Embeddings on Parallel Corpora (https://arxiv.org/abs/2101.08231)

Paper 2: Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study (https://arxiv.org/abs/2009.14304)

Additional paper: Multilingual Alignment of Contextual Word Representations (https://arxiv.org/abs/2002.03518)

Additional paper: Multilingual BERT Post-Pretraining Alignment (https://arxiv.org/abs/2010.12547)

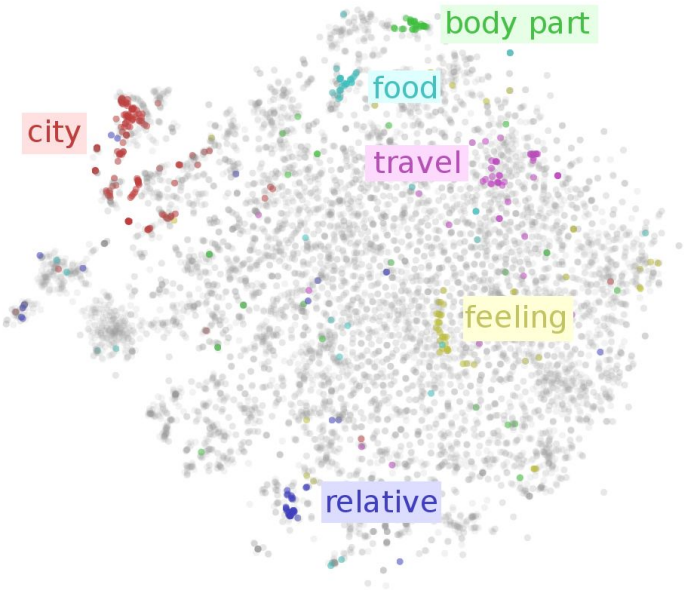
Additional paper: Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing (https://arxiv.org/abs/1909.06775)

Additional paper: Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing (https://arxiv.org/abs/1902.09492)

What are Embeddings?

Rome Paris word V

```
Rome   = [1, 0, 0, 0, 0, 0, …, 0]
Paris  = [0, 1, 0, 0, 0, 0, …, 0]
Italy  = [0, 0, 1, 0, 0, 0, …, 0]
France = [0, 0, 0, 1, 0, 0, …, 0]
```

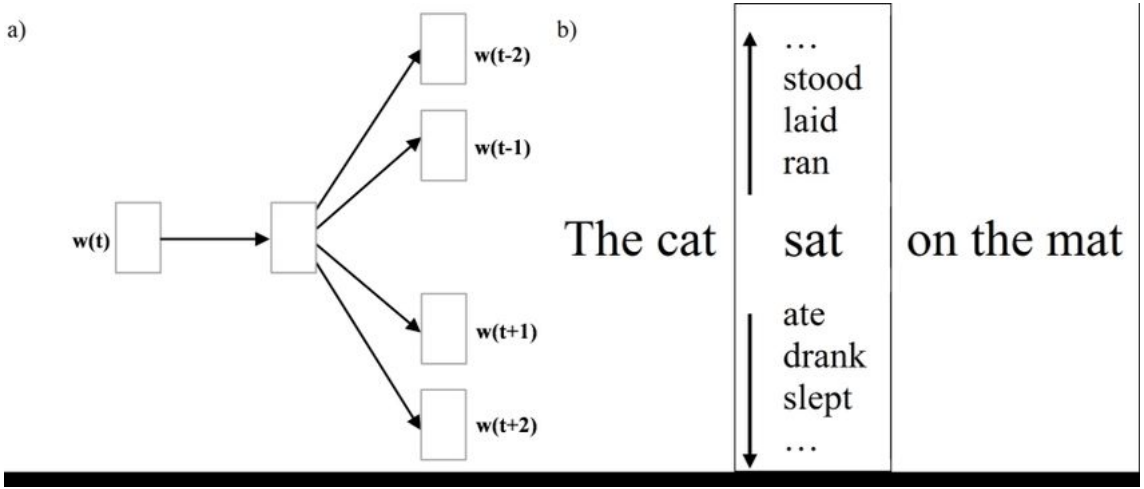


https://medium.com/intelligentmachines/word-embedding-and-one-hot-encoding-ad17b4bbe111

https://ruder.io/word-embeddings-1/

Distributional Representations

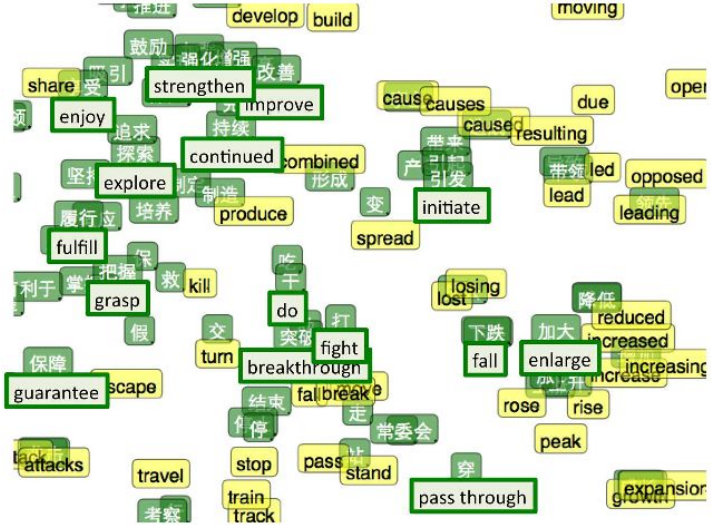
“You shall know a word by the company it keeps” --John Rupert Firth



https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c

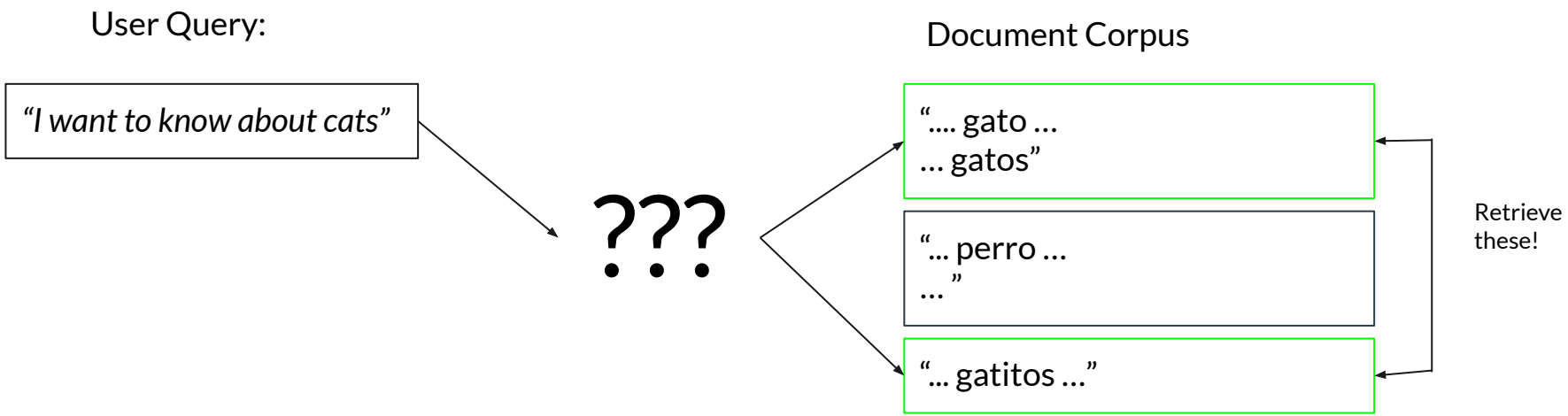
Multilingual Embeddings

- Embeddings for multiple languages in the same semantic space
- Can be jointly trained, or can “align” pretrained monolingual embedding spaces



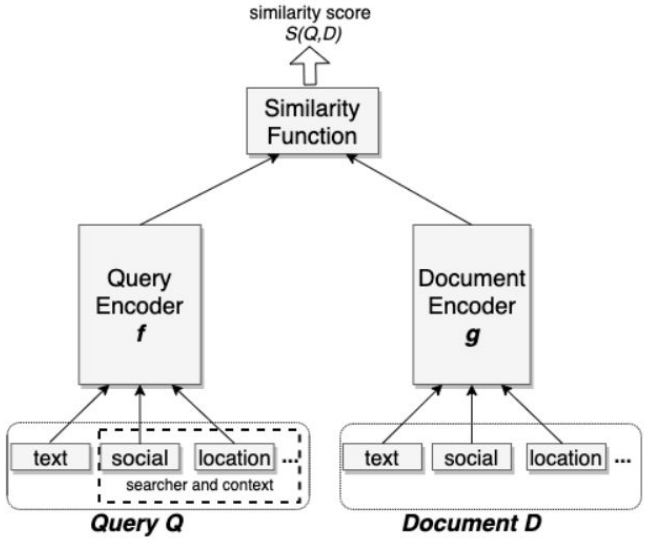
https://ruder.io/cross-lingual-embeddings/

Cross-lingual Information Retrieval (CLIR)



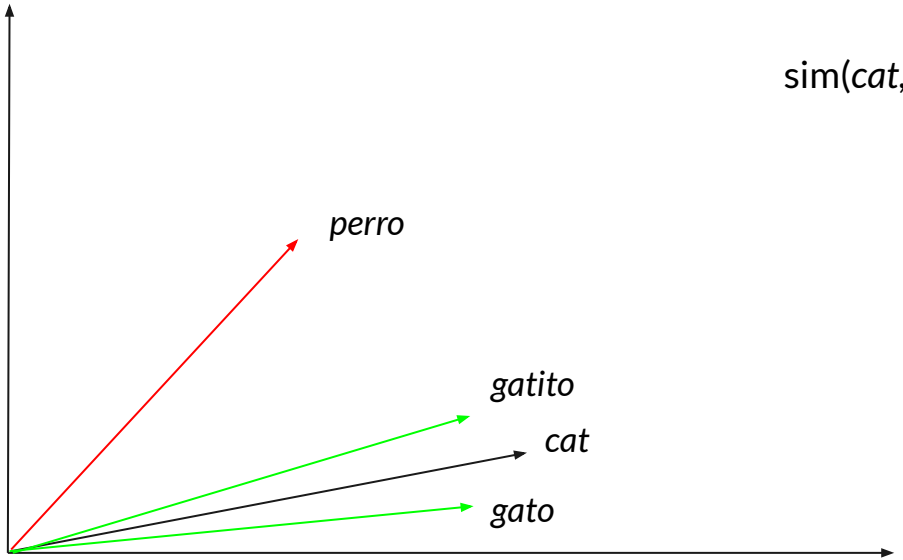
Cross-Lingual Information Retrieval

- Translating documents or query a common approach
  - Not doable on command if document corpus large enough
  - Documents/query in low-resource language->MT less powerful
- Monolingual approach to IR: encode doc and query, then calculate similarity, e.g. cosine similarity



https://research.fb.com/wp-content/uploads/2020/08/Embedding-based-Retrieval-in-Facebook-Search.pdf

Multilingual Embeddings for CLIR



sim(*cat*, *gato*) >> sim(*cat*, *perro*)

Neural Machine Translation: Word Translation

WORD TRANSLATION WITHOUT PARALLEL DATA

Alexis Conneau[*†‡], **Guillaume Lample**[*†§],
Marc'Aurelio Ranzato[†], **Ludovic Denoyer**[§], **Hervé Jégou**[†]
{aconneau,glample,ranzato,rvj}@fb.com
ludovic.denoyer@upmc.fr

ABSTRACT

State-of-the-art methods for learning cross-lingual word embeddings have relied on bilingual dictionaries or parallel corpora. Recent studies showed that the need for parallel data supervision can be alleviated with character-level information. While these methods showed encouraging results, they are not on par with their supervised counterparts and are limited to pairs of languages sharing a common alphabet. In this work, we show that we can build a bilingual dictionary between two languages without using any parallel corpora, by aligning monolingual word embedding spaces in an unsupervised way. Without using any character information, our model even outperforms existing supervised methods on cross-lingual tasks for some language pairs. Our experiments demonstrate that our method works very well also for distant language pairs, like English-Russian or English-Chinese. We finally describe experiments on the English-Esperanto low-resource language pair, on which there only exists a limited amount of parallel data, to show the potential impact of our method in fully unsupervised machine translation. Our code, embeddings and dictionaries are publicly available[1].

	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision (WaCky)						
Mikolov et al. (2013b) †	33.8	48.3	53.9	24.9	41.0	47.4
Dinu et al. (2015)†	38.5	56.4	63.9	24.6	45.4	54.1
CCA†	36.1	52.7	58.1	31.0	49.9	57.0
Artetxe et al. (2017)	39.7	54.7	60.5	33.8	52.4	59.1
Smith et al. (2017)†	43.1	60.7	66.4	38.0	58.5	63.6
Procrustes - CSLS	44.9	61.8	66.6	38.5	57.2	63.0
Methods without cross-lingual supervision (WaCky)						
Adv - Refine - CSLS	45.1	60.7	65.1	38.3	57.8	62.8
Methods with cross-lingual supervision (Wiki)						
Procrustes - CSLS	63.7	78.6	81.1	56.3	76.2	80.6
Methods without cross-lingual supervision (Wiki)						
Adv - Refine - CSLS	**66.2**	**80.4**	**83.4**	**58.7**	**76.5**	**80.9**

Conneau and Lample, ICLR 2018

Cross-lingual Embeddings for NMT

- Artetxe et al, ICLR 2018
- Pretrained monolingual embeddings can be used for NMT in lieu of large parallel corpora, and induce cross-lingual alignment using smaller parallel data

UNSUPERVISED NEURAL MACHINE TRANSLATION

Mikel Artetxe, Gorka Labaka & Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

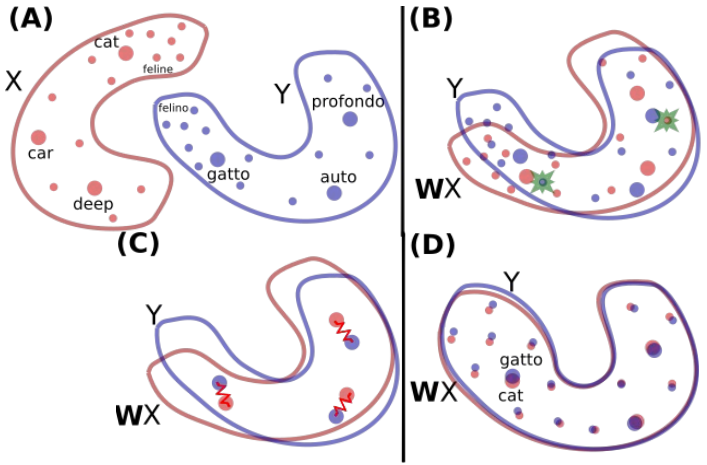
Kyunghyun Cho
New York University
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

ABSTRACT

In spite of the recent success of neural machine translation (NMT) in standard benchmarks, the lack of large parallel corpora poses a major practical problem for many language pairs. There have been several proposals to alleviate this issue with, for instance, triangulation and semi-supervised learning techniques, but they still require a strong cross-lingual signal. In this work, we completely remove the need of parallel data and propose a novel method to train an NMT system in a completely unsupervised manner, relying on nothing but monolingual corpora. Our model builds upon the recent work on unsupervised embedding mappings, and consists of a slightly modified attentional encoder-decoder model that can be trained on monolingual corpora alone using a combination of denoising and back-translation. Despite the simplicity of the approach, our system obtains 15.56 and 10.21 BLEU points in WMT 2014 French → English and German → English translation. The model can also profit from small parallel corpora, and attains 21.81 and 15.24 points when combined with 100,000 parallel sentences, respectively. Our implementation is released as an open source project[1].

What is embedding alignment?

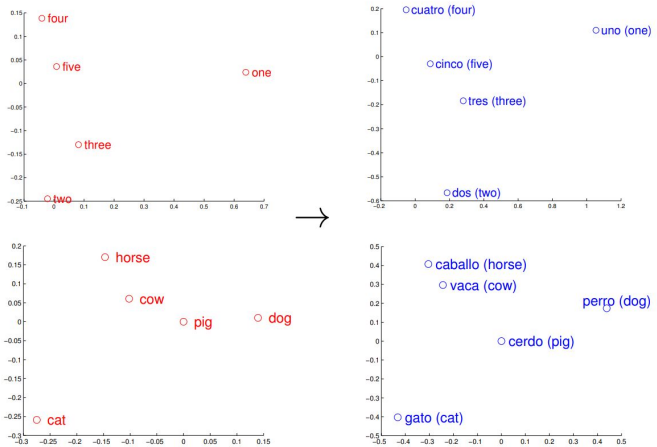
- Decreasing distance between 2 semantic spaces
- Creation of “multilingual” embeddings



https://github.com/facebookresearch/MUSE

(Static) Alignment: Rotational

- First proposed by [Mikolov et al. 2013]
- Learn a linear transformation W mapping one embedding space onto another



Rotational Alignment

- Artetxe et al., 2016
- Minimize sum of Euclidean distances of embeddings of word pairs
- Orthogonality constraint for computational efficiency and to preserve dot products in XW from X … is this a good assumption?

$$\arg\min_{W} \sum_{i} \|X_{i*}W - Z_{i*}\|^2$$

Multilingual BERT

- Trained on 104 languages
- Token embeddings based on sentence context as well as the actual token
- 12 embedding layers

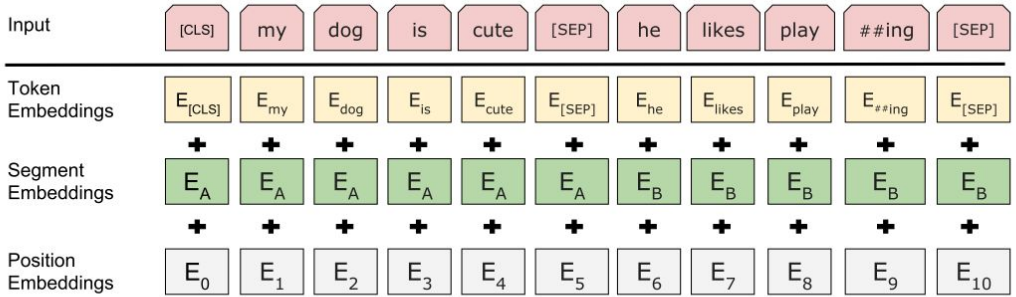
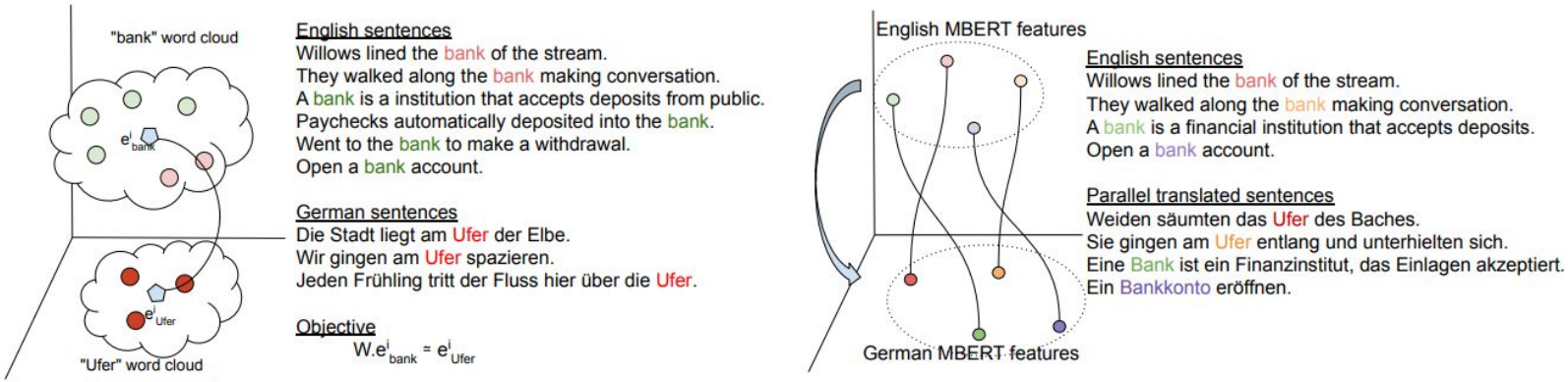


Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Contextual embeddings

- From static embeddings to “word clouds” -- allows for sense disambiguation



Kulshreshtha et al, ACL 2020

Contextual Embedding Alignment

- Do embedding spaces still have similar geometries?

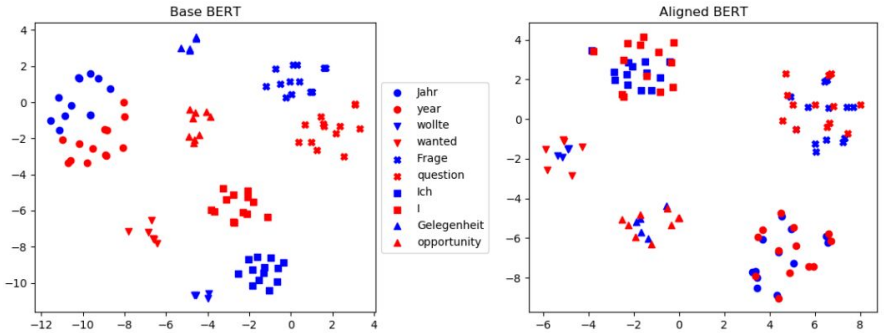


Figure 1: t-SNE (Maaten & Hinton, 2008) visualization of the embedding space of multilingual BERT for English-German word pairs (left: pre-alignment, right: post-alignment). Each point is a different instance of the word in the Europarl corpus. This figure suggests that BERT begins already somewhat aligned out-of-the-box but becomes much more aligned after our proposed procedure.

Cao et al. 2020

Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study (Paper 2)

- Kulshreshtha et al, ACL 2020

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

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

Jos Luis Redondo-Garca 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.

Fine-tuning Alignment

- Rather than rotating entire semantic space of language at once, adjust mBERT model parameters to make embeddings of word pairs closer to each other

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

- Sum this loss over all word pairs in corpus!

Alignment Objective

- For each word pair {m, m^*} in the corpus, find L2 distance between their contextual embeddings
- Corresponding word embeddings being far from each other is undesirable--so make them closer together!

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

Regularization Objective

- Recall that embedding spaces have some geometry and structure (Orthogonality assumption)
- Just blindly adjusting embeddings may destroy this structure: want to preserve locations as much as possible
- Make a copy of the mBERT model with frozen parameters and compare new embeddings to the originals

$$L^i_{regularise} = min \sum_m \| e^i_{s_m} - \mathbf{e}^i_{s_m} \|$$

Improving Rotation

- Applying a transformation naively is not the best approach--how to refine it?

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

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

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^*}:

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

Results

- Improve over non-aligned mBERT models!
- Rotation better for NER, Fine-tuning better for Semantic Slot Filling

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**

Word Alignment by Fine-tuning Embeddings on Parallel Corpora (Paper 1)

- Unique approach: using mBERT to extract word alignments
- Fine-tuned on that task using a combination of objectives
- “awesome-align”

Word Alignment by Fine-tuning Embeddings on Parallel Corpora

Zi-Yi Dou, Graham Neubig
Language Technologies Institute, Carnegie Mellon University
{zdou,gneubig}@cs.cmu.edu

Abstract

Word alignment over parallel corpora has a wide variety of applications, including learning translation lexicons, cross-lingual transfer of language processing tools, and automatic evaluation or analysis of translation outputs. The great majority of past work on word alignment has worked by performing unsupervised learning on parallel text. Recently, however, other work has demonstrated that pre-trained contextualized word embeddings derived from multilingually trained language models (LMs) prove an attractive alternative, achieving competitive results on the word alignment task even in the absence of explicit training on parallel data. In this paper, we examine methods to marry the two approaches: leveraging pre-trained LMs but fine-tuning them on parallel text with objectives designed to improve alignment quality, and proposing methods to effectively extract alignments from these fine-tuned models. We perform experiments on five language pairs and demonstrate that our model can consistently outperform previous state-of-the-art models of all varieties. In addition, we demonstrate that we are able to train multilingual word aligners that can obtain robust performance on different language pairs. Our aligner, **AWESOME** (**A**ligning **W**ord **E**mbedding **S**paces **O**f **M**ultilingual **E**ncoders), with pre-trained models is available at https://github.com/neulab/awesome-align.

1 Introduction

Word alignment is a useful tool to tackle a variety of natural language processing (NLP) tasks, including learning translation lexicons (Ammar et al., 2016; Cao et al., 2019), cross-lingual transfer of language processing tools (Yarowsky et al., 2001; Padó and Lapata, 2009; Tiedemann, 2014; Agić et al., 2016; Mayhew et al., 2017; Nicolai and Yarowsky, 2019), semantic parsing (Herzig and Berant, 2018) and

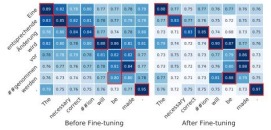


Figure 1: Cosine similarities between subword representations in a parallel sentence pair before and after fine-tuning. Red boxes indicate the gold alignments.

speech recognition (Xu et al., 2019). In particular, word alignment plays a crucial role in many machine translation (MT) related methods, including guiding learned attention (Liu et al., 2016), incorporating lexicons during decoding (Arthur et al., 2016), domain adaptation (Hu et al., 2019), unsupervised MT (Ren et al., 2020) and automatic evaluation or analysis of translation models (Bau et al., 2018; Stanovsky et al., 2019; Neubig et al., 2019; Wang et al., 2020). However, with neural networks advancing the state of the arts in almost every field of NLP, tools developed based on the 30-year-old IBM word-based translation models (Brown et al., 1993), such as GIZA++ (Och and Ney, 2003) or fast-align (Dyer et al., 2013), remain popular choices for word alignment tasks.

One alternative to using statistical word-based translation models to learn alignments would be to instead train state-of-the-art neural machine translation (NMT) models on parallel corpora, and extract alignments therefrom, as examined by Luong et al. (2015); Garg et al. (2019); Zenkel et al. (2020). However, these methods have two disadvantages (also shared with more traditional alignment methods): (1) they are directional and the source and target side are treated differently and (2) they cannot easily take advantage of large-scale contextualized

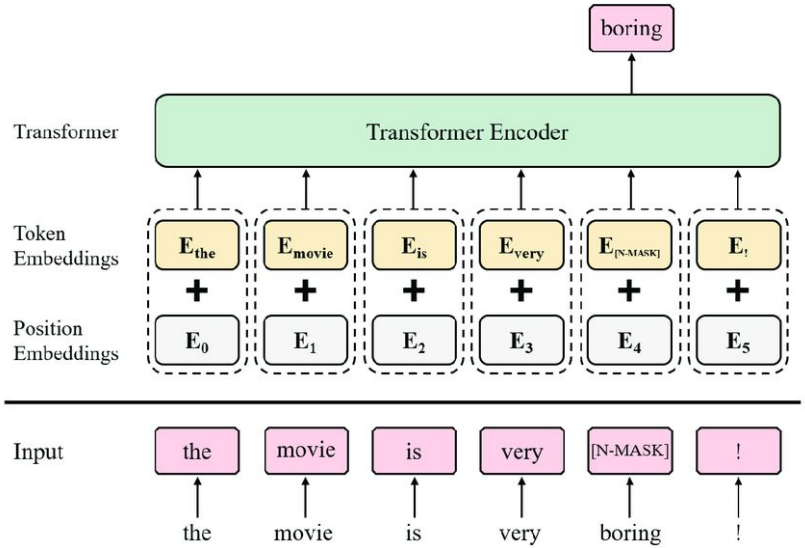
Training Objective

- Combination of language modeling objectives and word alignment-specific objectives

$$L = L_{MLM} + L_{TLM} + L_{SO} + L_{PSI} + \beta L_{CO}$$

Masked Language Modeling (MLM)

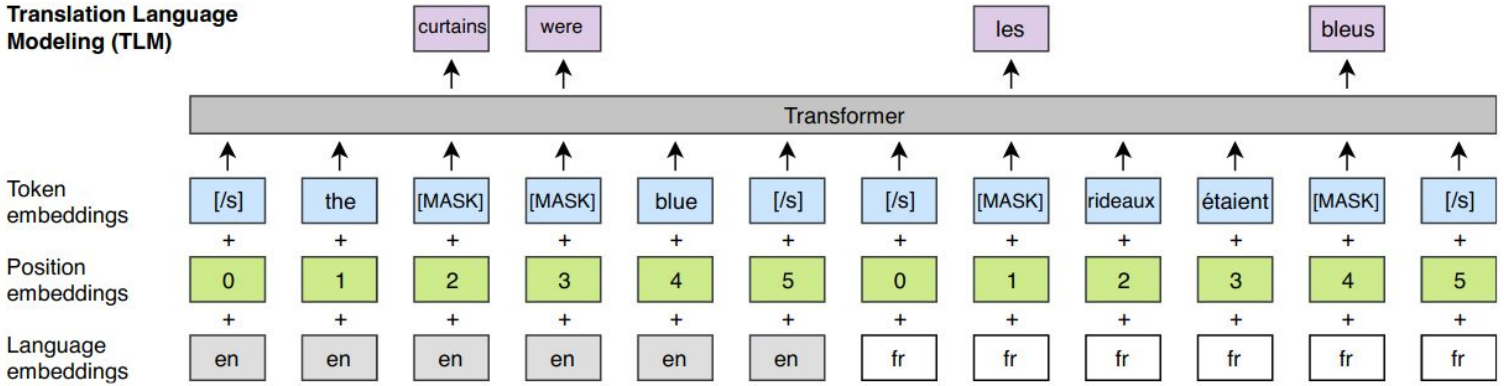
$$L_{MLM} = \log p(\mathbf{x}|\mathbf{x}^{mask}) + \log p(\mathbf{y}|\mathbf{y}^{mask})$$



https://www.researchgate.net/figure/Model-structure-of-the-label-masked-language-model-N-MASK-is-a-mask-token-containing_fig2_337187647

Translation Language Modeling (TLM)

$$L_{TLM} = \log p([\mathbf{x};\mathbf{y}]|[\mathbf{x}^{mask};\mathbf{y}^{mask}]) + \log p([\mathbf{y};\mathbf{x}]|[\mathbf{y}^{mask};\mathbf{x}^{mask}])$$



https://towardsdatascience.com/xlm-cross-lingual-language-model-33c1fd1adf82

Parallel Sentence Identification

- Make parallel sentence pair representations closer to each other

$$L_{PSI} = l \log s(\mathbf{x}', \mathbf{y}') + (1 - l) \log(1 - s(\mathbf{x}', \mathbf{y}'))$$

Self-training Objective

- A = extracted alignments, S = normalized cosine similarity matrix

$$L_{SO} = \sum_{i,j} A_{ij} \frac{1}{2} \left(\frac{S_{\mathbf{xy}_{ij}}}{n} + \frac{S^{\mathrm{T}}_{\mathbf{yx}_{ij}}}{m} \right)$$

Consistency Optimization

- Force symmetry between word alignments in both directions

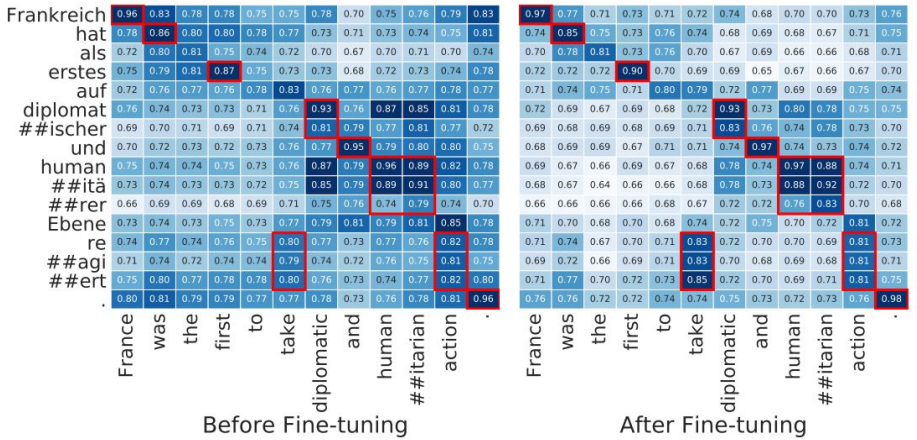
$$L_{CO} = \frac{\text{trace}(S_{\mathbf{xy}}^{\mathrm{T}} S_{\mathbf{yx}})}{\min(m, n)}$$

Cosine similarity between aligned words improves!

Outperforms other baseline word aligners as well

Model	Setting	De-En	Fr-En	Ro-En	Ja-En	Zh-En
Baseline						
SimAlign	*w/o fine-tuning*	18.8	7.6	27.2	46.6	21.6
fast_align	*bilingual*	27.0	10.5	32.1	51.1	38.1
eflomal	*bilingual*	22.6	8.2	25.1	47.5	28.7
GIZA++	*bilingual*	20.6	5.9	26.4	48.0	35.1
Zenkel et al. (2020)	*bilingual*	16.0	5.0	23.4	-	-
Chen et al. (2020)	*bilingual*	15.4	4.7	21.2	-	-
Ours						
α-entmax	*w/o fine-tuning*	18.1	5.6	29.0	46.3	18.4
	bilingual	16.1	***4.1***	23.4	38.6	15.4
	multilingual ($\beta = 0$)	15.4	***4.1***	22.9	***37.4***	**13.9**
	multilingual ($\beta = 1$)	***15.0***	4.5	**20.8**	38.7	14.5
	zero-shot	16.0	4.3	28.4	44.0	**13.9**
softmax	*w/o fine-tuning*	17.4	5.6	27.9	45.6	18.1
	bilingual	15.6	**4.4**	23.0	38.4	15.3
	multilingual ($\beta = 0$)	15.3	**4.4**	22.6	**37.9**	***13.6***
	multilingual ($\beta = 1$)	**15.1**	4.5	***20.7***	38.4	14.5
	zero-shot	15.7	4.6	27.2	43.7	14.0

Table 2: Performance (AER) of our models in bilingual, multilingual and zero-shot settings. The best scores for each alignment extraction method are in **bold** and the overall best scores are in ***italicized bold***.



Multilingual Alignment of Contextual Word Representations

- Origin of L2 fine-tuning objective used in Kulshreshtha et al.
- Unique due to lack of sentence-level or language modeling objectives

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

Published as a conference paper at ICLR 2020

MULTILINGUAL ALIGNMENT OF CONTEXTUAL WORD REPRESENTATIONS

Steven Cao, Nikita Kitaev & Dan Klein
Computer Science Division
University of California, Berkeley
{stevencao,kitaev,klein}@berkeley.edu

ABSTRACT

We propose procedures for evaluating and strengthening contextual embedding alignment and show that they are useful in analyzing and improving multilingual BERT. In particular, after our proposed alignment procedure, BERT exhibits significantly improved zero-shot performance on XNLI compared to the base model, remarkably matching pseudo-fully-supervised translate-train models for Bulgarian and Greek. Further, to measure the degree of alignment, we introduce a contextual version of word retrieval and show that it correlates well with downstream zero-shot transfer. Using this word retrieval task, we also analyze BERT and find that it exhibits systematic deficiencies, e.g. worse alignment for open-class parts-of-speech and word pairs written in different scripts, that are corrected by the alignment procedure. These results support contextual alignment as a useful concept for understanding large multilingual pre-trained models.

1 INTRODUCTION

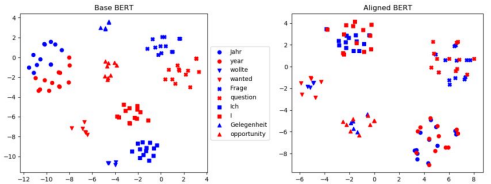


Figure 1: t-SNE (Maaten & Hinton, 2008) visualization of the embedding space of multilingual BERT for English-German word pairs (left: pre-alignment, right: post-alignment). Each point is a different instance of the word in the Europarl corpus. This figure suggests that BERT begins already somewhat aligned out-of-the-box but becomes much more aligned after our proposed procedure.

Embedding alignment was originally studied for word vectors with the goal of enabling cross-lingual transfer, where the embeddings for two languages are in alignment if word translations, e.g. *cat* and *Katze*, have similar representations (Mikolov et al., 2013a; Smith et al., 2017). Recently, large pre-trained models have largely subsumed word vectors based on their accuracy on downstream tasks, partly due to the fact that their word representations are context-dependent, allowing them to more richly capture the meaning of a word (Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018). Therefore, with the same goal of cross-lingual transfer but for these more complex models, we might consider contextual embedding alignment, where we observe whether word pairs within parallel sentences, e.g. *cat* in *"The cat sits"* and *Katze* in *"Die Katze sitzt,"* have similar representations.

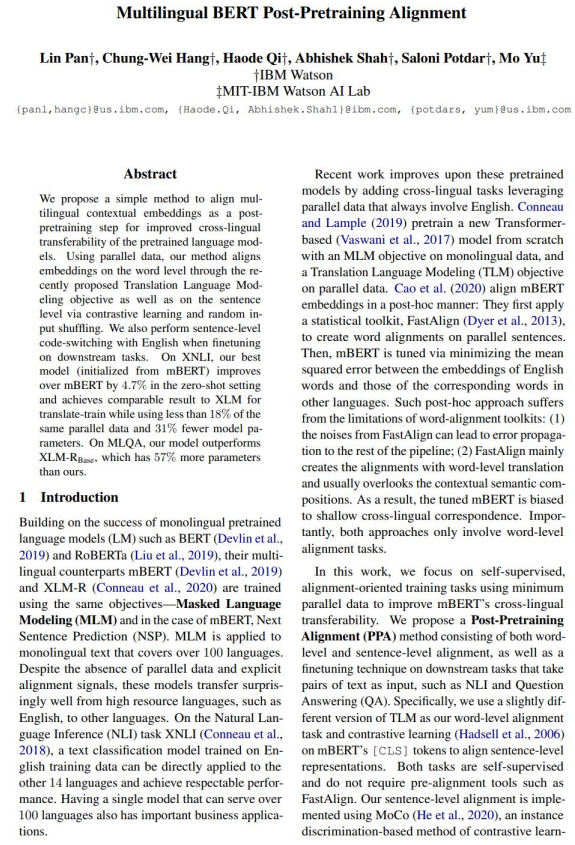
Results

- Word Retrieval task
- Rotation does not outperform fastText baseline, but fine-tuning does

	bg-en	de-en	el-en	es-en	fr-en	Average
Contextual						
Aligned fastText + sentence	44.0	46.4	42.0	48.6	44.5	45.1
Base BERT	19.5	26.1	13.9	32.5	28.3	24.1
Word-aligned BERT (rotation)	29.8	31.6	20.8	36.8	31.0	30.0
Word-aligned BERT (fine-tuned)	**50.7**	**51.3**	**49.8**	**51.0**	**48.6**	**50.3**
Non-Contextual						
Aligned fastText + sentence	61.3	**65.4**	61.6	**71.1**	64.8	64.8
Base BERT	29.1	37.0	22.3	46.5	41.8	35.3
Word-aligned BERT (rotation)	39.6	43.6	32.4	51.4	46.1	42.6
Word-aligned BERT (fine-tuned)	**62.8**	64.3	**67.5**	68.4	**66.3**	**65.9**

Multilingual BERT Post-Pretraining Alignment

- Pan et al. 2021



Multilingual BERT Post-Pretraining Alignment

Lin Pan†, Chung-Wei Hang†, Haode Qi†, Abhishek Shah†, Saloni Potdar†, Mo Yu‡
†IBM Watson
‡MIT-IBM Watson AI Lab
{panl,hangc}@us.ibm.com, {Haode.Qi, Abhishek.Shah1}@ibm.com, {potdars, yum}@us.ibm.com

Abstract

We propose a simple method to align multilingual contextual embeddings as a post-pretraining step for improved cross-lingual transferability of the pretrained language models. Using parallel data, our method aligns embeddings on the word level through the recently proposed Translation Language Modeling objective as well as on the sentence level via contrastive learning and random input shuffling. We also perform sentence-level code-switching with English when finetuning on downstream tasks. On XNLI, our best model (initialized from mBERT) improves over mBERT by 4.7% in the zero-shot setting and achieves comparable result to XLM for translate-train while using less than 18% of the same parallel data and 31% fewer model parameters. On MLQA, our model outperforms XLM-R_{Base}, which has 57% more parameters than ours.

1 Introduction

Building on the success of monolingual pretrained language models (LM) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), their multilingual counterparts mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are trained using the same objectives—**Masked Language Modeling (MLM)** and in the case of mBERT, Next Sentence Prediction (NSP). MLM is applied to monolingual text that covers over 100 languages. Despite the absence of parallel data and explicit alignment signals, these models transfer surprisingly well from high resource languages, such as English, to other languages. On the Natural Language Inference (NLI) task XNLI (Conneau et al., 2018), a text classification model trained on English training data can be directly applied to the other 14 languages and achieve respectable performance. Having a single model that can serve over 100 languages also has important business applications.

Recent work improves upon these pretrained models by adding cross-lingual tasks leveraging parallel data that always involve English. Conneau and Lample (2019) pretrain a new Transformer-based (Vaswani et al., 2017) model from scratch with an MLM objective on monolingual data, and a Translation Language Modeling (TLM) objective on parallel data. Cao et al. (2020) align mBERT embeddings in a post-hoc manner: They first apply a statistical toolkit, FastAlign (Dyer et al., 2013), to create word alignments on parallel sentences. Then, mBERT is tuned via minimizing the mean squared error between the embeddings of English words and those of the corresponding words in other languages. Such post-hoc approach suffers from the limitations of word-alignment toolkits: (1) the noises from FastAlign can lead to error propagation to the rest of the pipeline; (2) FastAlign mainly creates the alignments with word-level translation and usually overlooks the contextual semantic compositions. As a result, the tuned mBERT is biased to shallow cross-lingual correspondence. Importantly, both approaches only involve word-level alignment tasks.

In this work, we focus on self-supervised, alignment-oriented training tasks using minimum parallel data to improve mBERT's cross-lingual transferability. We propose a **Post-Pretraining Alignment (PPA)** method consisting of both word-level and sentence-level alignment, as well as a finetuning technique on downstream tasks that take pairs of text as input, such as NLI and Question Answering (QA). Specifically, we use a slightly different version of TLM as our word-level alignment task and contrastive learning (Hadsell et al., 2006) on mBERT's `[CLS]` tokens to align sentence-level representations. Both tasks are self-supervised and do not require pre-alignment tools such as FastAlign. Our sentence-level alignment is implemented using MoCo (He et al., 2020), an instance discrimination-based method of contrastive learn-

Fine-tuning Objective

- L_{MoCo} a contrastive objective to keep sentence representations with same meaning “close”, and pairs with different meaning “far”
- L_{TLM} a Translation Modeling Objective

$$\mathcal{L} = \mathcal{L}_{\text{MoCo}} + \mathcal{L}_{\text{TLM}}$$

- Recall from awesome-align: this is almost a subset of these objectives!

$$L = L_{MLM} + L_{TLM} + L_{SO} + L_{PSI} + \beta L_{CO},$$

Results

- XNLI dataset
- XLM a multilingual model with more parameters

Model	en	fr	es	de	bg	ar	zh	hi	avg
Zero-shot cross-lingual transfer									
mBERT (Devlin et al., 2019)	81.4	-	74.3	70.5	-	62.1	63.8	-	-
mBERT from (Hu et al., 2020)	80.8	73.4	73.5	70.0	68.0	64.3	67.8	58.9	69.6
Cao et al. (2020)	80.1	74.5	75.5	73.1	73.4	-	-	-	-
Artetxe and Schwenk (2019)	73.9	71.9	72.9	72.6	74.2	**71.4**	71.4	65.5	71.7
Ours (250k)	82.4	75.5	76.2	73.3	**74.6**	68.2	71.7	62.8	73.1
Ours (600k)	82.4	**76.7**	76.4	74.0	74.1	69.1	72.3	66.9	74.0
Ours (2M)	**82.8**	76.6	**76.7**	**74.2**	73.8	70.3	**72.8**	**66.9**	**74.3**
XLM (MLM)	83.2	76.5	76.3	74.2	74.0	68.5	71.9	65.7	73.8
XLM (MLM + TLM)	85.0	78.7	78.9	77.8	77.4	73.1	76.5	69.6	77.1
Translate-train									
mBERT (Devlin et al., 2019)	81.9	-	77.8	75.9	-	70.7	76.6	-	-
mBERT from (Wu and Dredze, 2019)	82.1	76.9	78.5	74.8	75.4	70.8	76.2	65.3	75.0
Ours (250k)	82.4	78.8	79.0	**78.7**	78.4	74.0	77.9	69.6	77.4
Ours (600k)	82.4	79.7	79.7	77.9	**79.0**	75.2	77.8	71.5	77.9
Ours (2M)	**82.8**	**79.7**	**80.6**	78.6	78.8	**75.2**	**78.0**	**72.0**	**78.2**
XLM (Conneau and Lample, 2019)	85.0	80.2	80.8	80.3	79.3	76.5	78.6	72.3	79.1

Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing

- Schuster et al. 2019

Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing

Tal Schuster[*,1], **Ori Ram**[*,2], **Regina Barzilay**[1], **Amir Globerson**[2]

[1]Computer Science and Artificial Intelligence Lab, MIT
[2]Tel Aviv University
{tals, regina}@csail.mit.edu, {ori.ram, gamir}@cs.tau.ac.il

Abstract

We introduce a novel method for multilingual transfer that utilizes deep contextual embeddings, pretrained in an unsupervised fashion. While contextual embeddings have been shown to yield richer representations of meaning compared to their static counterparts, aligning them poses a challenge due to their dynamic nature. To this end, we construct context-independent variants of the original monolingual spaces and utilize their mapping to derive an alignment for the context-dependent spaces. This mapping readily supports processing of a target language, improving transfer by context-aware embeddings. Our experimental results demonstrate the effectiveness of this approach for zero-shot and few-shot learning of dependency parsing. Specifically, our method consistently outperforms the previous state-of-the-art on 6 tested languages, yielding an improvement of 6.8 LAS points on average.[1]

1 Introduction

Multilingual embedding spaces have been demonstrated to be a promising means for enabling cross-lingual transfer in many natural language processing tasks (e.g. Ammar et al. (2016); Lample et al. (2018)). Similar to how universal part-of-speech tags enabled parsing transfer across languages (Petrov et al., 2012), multilingual word embeddings further improve transfer capacity by enriching models with lexical information. Since this lexical representation is learned in an unsupervised fashion and thus can leverage large amounts of raw data, it can capture a more nuanced representation of meaning than unlexicalized transfer. Naturally, this enrichment is translated into improved transfer accuracy, especially in low-resource scenarios (Guo et al., 2015).

In this paper, we are moving further along this line and exploring the use of contextual word embeddings for multilingual transfer. By dynamically linking words to their various contexts, these embeddings provide a richer semantic and syntactic representation than traditional context-independent word embeddings (Peters et al., 2018). A straightforward way to utilize this richer representation is to directly apply existing transfer algorithms on the contextual embeddings instead of their static counterparts. In this case, however, each token pair is represented by many different vectors corresponding to its specific context. Even when supervision is available in the form of a dictionary, it is still unclear how to utilize this information for multiple contextual embeddings that correspond to a word translation pair.

In this paper, we propose a simple but effective mechanism for constructing a multilingual space of contextual embeddings. Instead of learning the alignment in the original, complex contextual space, we drive the mapping process using context-independent embedding anchors. We obtain these anchors by factorizing the contextual embedding space into context-independent and context-dependent parts. Operating at the anchor level not only compresses the space, but also enables us to utilize a word-level bilingual dictionary as a source of supervision, if available. Once the anchor-level alignment is learned, it can be readily applied to map the original spaces with contextual embeddings.

Clearly, the value of word embeddings depends on their quality, which is determined by the amount of raw data available for their training (Jiang et al., 2018). We are interested in expanding the above approach to the truly low-resource

[*] Equal contribution
[1]Code and models: https://github.com/TalSchuster/CrossLingualELMo.

“Anchors”

- Context-independent variants of contextual embeddings

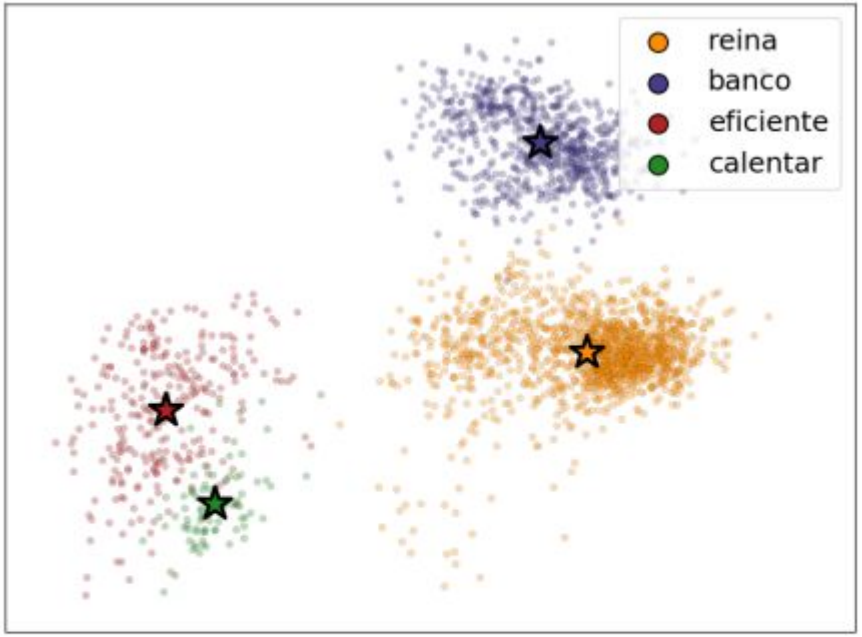


Figure 1: A two dimensional PCA showing examples of contextual representations for four Spanish words. Their corresponding anchors are presented as a star in the same color. (best viewed in color)

Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing

- Wang et al., 2019
- Rotation-based method
- **Word-sense preserving**
- Test both an SVD and Gradient Descent approach to compute transformation W
- Orthogonality constraint removed for GD approach, but still approximating SVD solution (orthogonal)

Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing

Yuxuan Wang[1], **Wanxiang Che**[1]*, **Jiang Guo**[2], **Yijia Liu**[1], **and Ting Liu**[1]
[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
[2]Computer Science and Artificial Intelligence Laboratory, MIT
{yxwang, car, yjliu, tliu}@ir.hit.edu.cn, jiang_guo@csail.mit.edu

Abstract

This paper investigates the problem of learning cross-lingual representations in a contextual space. We propose **C**ross-**L**ingual **B**ERT **T**ransformation (CLBT), a simple and efficient approach to generate cross-lingual contextualized word embeddings based on publicly available pre-trained BERT models (Devlin et al., 2018). In this approach, a linear transformation is learned from contextual word alignments to align the contextualized embeddings independently trained in different languages. We demonstrate the effectiveness of this approach on zero-shot cross-lingual transfer parsing. Experiments show that our embeddings substantially outperform the previous state-of-the-art that uses static embeddings. We further compare our approach with XLM (Lample and Conneau, 2019), a recently proposed cross-lingual language model trained with massive parallel data, and achieve highly competitive results. [1]

1 Introduction

One of the most promising directions for cross-lingual dependency parsing, which also remains a challenge, is to bridge the gap of lexical features. Prior works (Xiao and Guo, 2014; Guo et al., 2015) have shown that cross-lingual word embeddings are able to significantly improve the transfer performance compared to delexicalized models (McDonald et al., 2011, 2013). These cross-lingual word embeddings are *static* in the sense that they do not change with the context.[2]

Recently, contextualized word embeddings derived from large-scale pre-trained language models (McCann et al., 2017; Peters et al., 2017, 2018; Devlin et al., 2018) have demonstrated dramatic superiority over traditional static word embeddings, establishing new state-of-the-arts in various monolingual NLP tasks (Ilić et al., 2018; Schuster et al., 2018). The success has also been recognized in dependency parsing (Che et al., 2018). The great potential of these contextualized embeddings has inspired us to extend its power to cross-lingual scenarios.

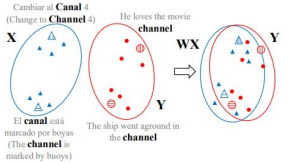


Figure 1: A toy illustration of the method, where contextualized embeddings of the word *canal* from Spanish is transformed to the semantic space of English.

Several recent works have been proposed to learn contextualized cross-lingual embeddings by training cross-lingual language models from scratch with parallel data as supervision, and has been demonstrated effective in several downstream tasks (Schuster et al., 2018; Mulcaire et al., 2019; Lample and Conneau, 2019). These methods are typically resource-demanding and time-consuming.[3] In this paper, we propose **C**ross-**L**ingual **B**ERT **T**ransformation (CLBT), a simple and efficient off-line approach that learns a linear transformation from contextual word alignments. With CLBT, contextualized embeddings

*Email corresponding
[1]Our code is released at https://github.com/WangYuxuan93/CLBT
[2]In this paper, we refer to these embeddings as *static* as opposed to *contextualized* ones.
[3]For instance, XLM was trained on 64 Volta GPUs (Lample and Conneau, 2019). While the time of training is not described in the paper, we may take the statistics from BERT as a reference, e.g., $\text{BERT}_{\text{BASE}}$ was trained on 4 Cloud TPUs for 4 days (Devlin et al., 2018).

Results

- Improvement using both SVD and GD methods
- Consistent improvement, but note performance decreases with language distance

Lan.	Static	Contextualized		
	FT-SVD	**mBERT**	**CLBT (SVD)**	**CLBT (GD)**
en	88.31	90.71	**91.03***	
de	59.31	63.41	**64.47***	62.14
da	68.81	70.57	71.60*	**71.66***
sv	73.49	70.09	73.33*	**75.95***
nl	60.11	**65.66**	65.45	63.86
fr	73.46	72.97	74.70*	**76.59***
it	76.23	79.02	**79.46**	78.98
es	66.91	65.43	67.14*	**68.33***
pt	67.98	67.11	69.12*	**69.25***
ro	52.11	46.40	55.14*	**55.84***
sk	56.98	50.76	59.46*	**59.92***
pl	58.59	63.10	65.37*	**65.80***
bg	66.68	**71.20**	70.26	70.75
sl	54.57	56.78	**57.42***	57.21*
cs	52.80	45.20	52.20*	**52.99***
fi	48.74	49.56	51.00*	**52.61***
et	44.40	46.64	47.79*	**48.52***
lv	49.59	45.11	48.59*	**49.78***
AVG.	60.63	60.53	63.09	**63.54**

Table 1: Results (LAS%) on test sets. Languages are split by language families with dashed lines. AVG. means the average of results from all target languages. Statistically significant differences between our methods and the mBERT model are marked with *, with p-value < 0.05 under McNemar's test.

Strengths

- Best results can be achieved with low parallel corpus size (only ~5k words or even smaller)
- Fast convergence

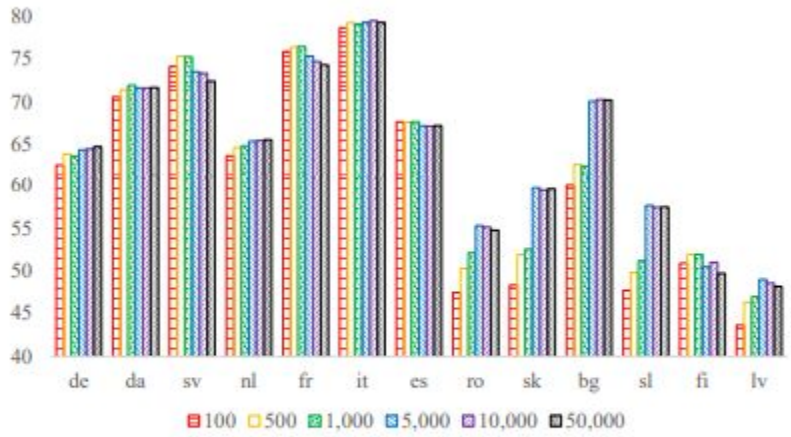


Figure 3: Effects of the amount of training data on different languages. (y-axis represents the LAS.)

Conclusion

- A number of different approaches to aligning contextual embeddings exist
  - Rotation of semantic spaces
  - Improvement of mBERT as a language model
  - “Post-hoc” embedding fine-tuning

Existing Tools

- MUSE
- Awesome-align
- CLBT

Visualization/Demo

- https://github.com/facebookresearch/MUSE/blob/main/demo.ipynb
- Awesome-align Demo

Discussion Questions

Discussion Questions

1. Are any of these techniques strictly better than others? Or do you expect it to be more task-specific?
2. Revisiting the orthogonality assumption--do you think this holds over all languages equally? Is it reasonable to assume that all languages have similar embedding space structures?
3. How might this affect low-resource languages differently?
4. How can we combat this disparity in performance?
    a. Are All Languages Created Equal in Multilingual BERT? (Wu and Dredze, ACL 2020)
5. Are there any tasks that (at least one alignment method) might not help to improve performance on? Why would you expect this?