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, **AWE-SoME** (**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
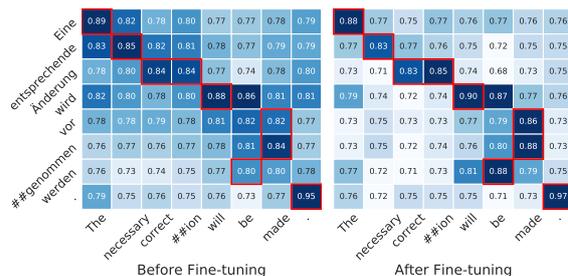

Figure 1: Cosine similarities between subword representations in a parallel sentence pair before and after fine-tuning. Red boxes indicate the gold alignments.

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

word embeddings derived from language models (LMs) multilingually trained on monolingual corpora (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020), which have proven useful in other cross-lingual transfer settings (Libovickỳ et al., 2019; Hu et al., 2020b). In the field of word alignment, Sabet et al. (2020) have recently proposed methods to align words using multilingual contextualized embeddings and achieve good performance even in the absence of explicit training on parallel data, suggesting that these are an attractive alternative for neural word alignment.

In this paper, we investigate if we can combine the best of the two lines of approaches. Concretely, we leverage pre-trained LMs and fine-tune them on parallel text with not only LM-based objectives, but also unsupervised objectives over the parallel corpus designed to improve alignment quality. Specifically, we propose a self-training objective, which encourages aligned words to have further closer contextualized representations, and a parallel sentence identification objective, which enables the model to bring parallel sentences' representations closer to each other. In addition, we propose to effectively extract alignments from these fine-tuned models using probability thresholding or optimal transport.

We perform experiments on five different language pairs and demonstrate that our model can achieve state-of-the-art performance on all of them. In analysis, we find that these approaches also generate more aligned contextualized representations after fine-tuning (see Figure 1 as an example) and we can incorporate supervised signals within our paradigm. Importantly, we show that it is possible to train multilingual word aligners that can obtain robust performance even in zero-shot settings, making them a valuable tool that can be used out-of-the-box with good performance over a wide variety of language pairs.

2 Methods

Formally, the task of word alignment can be defined as: given a sentence $\mathbf{x} = \langle x_1, \cdots, x_n \rangle$ in the source language and its corresponding parallel sentence $\mathbf{y} = \langle y_1, \cdots, y_m \rangle$ in the target language, a word aligner needs to find a set of pairs of source and target words:

$$A = \{\langle x_i, y_j \rangle : x_i \in \mathbf{x}, y_j \in \mathbf{y}\},$$

where for each word pair $\langle x_i, y_j \rangle$, x_i and y_j are semantically similar to each other within the context of the sentence.

In the following paragraphs, we will first illustrate how we extract alignments from contextualized word embeddings, then describe our objectives designed to improve alignment quality.

2.1 Extracting Alignments from Embeddings

Contextualized word embedding models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) represent words using continuous vectors calculated in context, and have achieved impressive performance on a diverse array of NLP tasks. Multilingually trained word embedding models such as multilingual BERT can generate contextualized embeddings across different languages. These models can be used to extract contextualized word embeddings $h_{\mathbf{x}} = \langle h_{x_1}, \cdots, h_{x_n} \rangle$ and $h_{\mathbf{y}} = \langle h_{y_1}, \cdots, h_{y_m} \rangle$ for each pair of parallel sentences $\mathbf{x}$ and $\mathbf{y}$. Specifically, this is done by extracting the hidden states of the i-th layer of the model, where i is an empirically-chosen hyper-parameter. Given these contextualized word embeddings, we propose two methods to calculate unidirectional alignment scores based on probability simplexes and optimal transport. We then turn these alignment scores into alignment matrices and reconcile alignments in the forward and backward directions.

Probability Thresholding. In this method, for each word in the source/target sentence, we calculate a value on the probability simplex for each word in the aligned target/source sentence, and then select all values that exceed a particular threshold as "aligned" words. Concretely, taking inspiration from attention mechanisms (Bahdanau et al., 2015; Vaswani et al., 2017), we take the contextualized embeddings $h_{\mathbf{x}}$ and $h_{\mathbf{y}}$ and compute the dot products between them and get the similarity matrix:

$$S = h_{\mathbf{x}} h_{\mathbf{y}}^T.$$

Then, we apply a normalization function $\mathcal{N}$ to convert the similarity matrix into values on the probability simplex $S_{\mathbf{xy}} = \mathcal{N}(S)$, and treat $S_{\mathbf{xy}}$ as the source-to-target alignment matrix. In this paper, we propose to use *softmax* and a sparse variant α-entmax (Peters et al., 2019) to do the normalization. Compared with the *softmax* function, α-entmax can produce sparse alignments for any $\alpha > 1$ and assign non-zero probability to a short
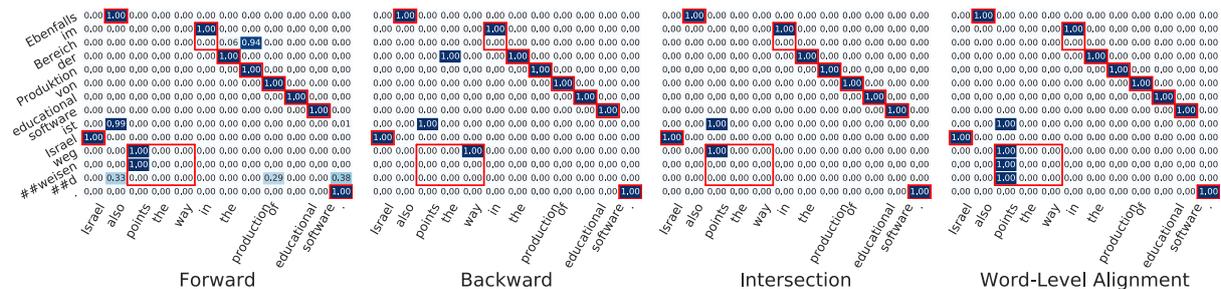
Figure 2: Extracting word alignments from multilingual BERT using probability thresholding (*softmax*). Red boxes denote the gold alignments.

list of plausible word pairs, where a higher α will lead to a more sparse alignment.

Optimal Transport. The goal of optimal transport (Monge, 1781; Cuturi, 2013) is to find a mapping that moves probability from one distribution to another, which can be used to find an optimal matching of similar words between two sequences (Kusner et al., 2015). Formally, in a discrete optimal transport problem, we are given two point sets $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ associated with their probability distributions $p_\mathbf{x}$ and $p_\mathbf{y}$ where $\sum_i p_{x_i} = 1$ and $\sum_j p_{y_j} = 1$. Also, a function $C(x_i, y_j)$ defines the cost of moving point x_i to y_j. The goal of optimal transport is to find a mapping that moves probability mass from $\{x_i\}_{i=1}^n$ to $\{y_j\}_{j=1}^m$ and the total cost of moving the mass between points is minimized. In other words, it finds the transition matrix $S_\mathbf{xy}$ that minimizes:

$$\sum_{i,j} C(x_i, y_j) S_{\mathbf{xy}ij}, \quad (1)$$

where $S_\mathbf{xy} \mathbf{1}_m = p_\mathbf{x}$ and $S_\mathbf{xy}^T \mathbf{1}_n = p_\mathbf{y}$. The resulting transition matrix is self-normalized and sparse (Swanson et al., 2020), making it appealing alternative towards extracting alignments from word embeddings.

In this paper, we propose to adapt optimal transport techniques to the task of word alignment. Concretely, we treat the parallel sentences $\mathbf{x}$ and $\mathbf{y}$ as two point sets and assume each word is uniformly distributed. The cost function is obtained by computing the pairwise distance (e.g. cosine distance) between $h_\mathbf{x}$ and $h_\mathbf{y}$, and all the distance values are scaled to [0, 1] with min-max normalization. The optimal transition matrix $S_\mathbf{xy}$ to Equation 1 can be calculated using the Sinkhorn-Knopp matrix scaling algorithm (Sinkhorn and Knopp, 1967). If the value of $S_{\mathbf{xy}ij}$ is high, x_i and y_j are likely to have

similar semantics and values that exceed a particular threshold will be considered as "aligned".

Extracting Bidirectional Alignments. After we obtain both the source-to-target and target-to-source alignment probability matrices $S_\mathbf{xy}$ and $S_\mathbf{yx}$ using the previous methods, we can deduce the final alignment matrix by taking the intersection of the two matrices:

$$A = (S_\mathbf{xy} > c) * (S_\mathbf{yx}^T > c),$$

where c is a threshold and $A_{ij} = 1$ means x_i and y_j are aligned.

Note that growing heuristics such as *grow-diag-final* (Och and Ney, 2000; Koehn et al., 2005) that are popular in statistical word aligners can also be applied in our alignment extraction algorithms, and we will demonstrate the effect of these heuristics in the experiment section.

Handling Subwords. Subword segmentation techniques (Sennrich et al., 2016; Kudo and Richardson, 2018) are widely used in training LMs, thus the above alignment extraction methods can only produce alignments on the subword level. To convert them to word alignments, we follow previous work (Sabet et al., 2020; Zenkel et al., 2020) and consider two words to be aligned if any of their subwords are aligned. Figure 2 shows a concrete example of how we extract word-level alignments from a pre-trained embedding model.

2.2 Fine-tuning Contextualized Embeddings for Word Alignment

While language models can be used to produce reasonable word alignments even without any fine-tuning (Sabet et al., 2020), we propose objectives that further improve their alignment ability if we have access to parallel data.

Masked Language Modeling (MLM). Guru-rangan et al. (2020) suggest that we can gain improvements in downstream tasks by further pre-training LMs on the task datasets. Therefore, we propose to fine-tune the LMs with a masked language modeling objective on both the source and target side of parallel corpora. Specifically, given a pair of parallel sentences $\mathbf{x}$ and $\mathbf{y}$, we choose 15% of the token positions randomly for both $\mathbf{x}$ and $\mathbf{y}$, and for each chosen token, we replace it with (1) the [MASK] token 80% of the time (2) a random token 10% of the time and (3) unchanged 10% of the time. The model is trained to reconstruct the original tokens given the masked sentences $\mathbf{x}^{mask}$ and $\mathbf{y}^{mask}$:

$$L_{MLM} = \log p(\mathbf{x}|\mathbf{x}^{mask}) + \log p(\mathbf{y}|\mathbf{y}^{mask}). \quad (2)$$

Translation Language Modeling (TLM). The MLM objective only requires monolingual data and the model cannot make direct connections between parallel sentences. To solve the issue, similarly to Lample and Conneau (2019), we concatenate parallel sentences $\mathbf{x}$ and $\mathbf{y}$ and perform MLM on the concatenated data. Compared with MLM, the translation language modeling (TLM) objective enable the model to align the source and target representations. Different from Lample and Conneau (2019), we feed source and target sentences twice in different orders instead of resetting the positions of target sentences:

$$L_{TLM} = \log p([\mathbf{x};\mathbf{y}]|[\mathbf{x}^{mask};\mathbf{y}^{mask}]) \\ + \log p([\mathbf{y};\mathbf{x}]|[\mathbf{y}^{mask};\mathbf{x}^{mask}]). \quad (3)$$

Self-training Objective (SO). We also propose a self-training objective for fine-tuning LMs which is similar to the EM algorithm used in the IBM models and the agreement constraints in Tamura et al. (2014). Specifically, at each training step, we first use our alignment extraction methods (described in Section 2.1) to extract the alignment A for $\mathbf{x}$ and $\mathbf{y}$, then maximize the following objective:

$$L_{SO} = \sum_{i,j} A_{ij}\frac{1}{2}(\frac{S_{\mathbf{xy}_{ij}}}{n} + \frac{S_{\mathbf{yx}_{ij}}}{m}). \quad (4)$$

Intuitively, this objective encourages words aligned in the first pass of alignment to have further closer contextualized representations. In addition, because of the intersection operation during extraction, the self-training objective can ideally reduce

	De-En	Fr-En	Ro-En	Ja-En	Zh-En
#Train Sents.	1.9M	1.1M	450K	444K	40K
#Test Sents.	508	447	248	582	450

Table 1: Statistics of datasets.

spurious alignments and encourage the source-to-target and target-to-source alignments to be symmetrical to each other by exploiting their agreement (Liang et al., 2006).

Parallel Sentence Identification (PSI). We also propose a contrastive parallel sentence identification loss that attempts to make parallel sentences more similar than mismatched sentence pairs (Liu and Sun, 2015; Legrand et al., 2016). This encourages the overall alignments of embeddings on both word and sentence level to be closer together. Concretely, we randomly select a pair of parallel or non-parallel sentences $\langle \mathbf{x}', \mathbf{y}' \rangle$ from the training data with equal probability. Then, the model is required to predict whether the two sampled sentences are parallel or not. The representation of the first [CLS] token is fed into a multi-layer perceptron to output a prediction score $s(\mathbf{x}', \mathbf{y}')$. Denoting the binary label as l, the objective function can be written as:

$$L_{PSI} = l \log s(\mathbf{x}', \mathbf{y}') + (1-l) \log(1-s(\mathbf{x}', \mathbf{y}')). \quad (5)$$

Consistency Optimization (CO). While the self-training objective can potentially improve the symmetricity between forward and backward alignments, following previous work on machine translation and multilingual representation learning (Cohn et al., 2016; Zhang et al., 2019; Hu et al., 2020a), we use an objective to explicitly encourage the consistency between the two alignment matrices. Specifically, we maximize the trace of $S_{\mathbf{xy}}^{\mathrm{T}} S_{\mathbf{yx}}$:

$$L_{CO} = -\frac{\mathrm{trace}(S_{\mathbf{xy}}^{\mathrm{T}} S_{\mathbf{yx}})}{\min(m,n)}. \quad (6)$$

Our Final Objective. In summary, our training objective is a combination of the proposed objectives and we train the model with them jointly at each training step:

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

where β is set to 0 or 1 in our experiments.

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						
	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
α-entmax	*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**
	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
softmax	*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***.

3 Experiments

In this section, we first present our main results, then conduct several ablation studies and analyses of our models.

3.1 Setup

Datasets. We perform experiments on five different language pairs, namely German-English (De-En), French-English (Fr-En), Romanian-English (Ro-En), Japanese-English (Ja-En) and Chinese-English (Zh-En). For the De-En, Fr-En, Ro-En datasets, we follow the experimental setting of previous work (Zenkel et al., 2019; Garg et al., 2019; Zenkel et al., 2020). The training and test data for Ro-En and Fr-En are provided by Mihalcea and Pedersen (2003). The Ro-En training data are also augmented by the Europarl v8 corpus (Koehn, 2005). For the De-En data, the Europarl v7 corpus is used as training data and the gold alignments are provided by Vilar et al. (2006). The Ja-En dataset is obtained from the Kyoto Free Translation Task (KFTT) word alignment data (Neubig, 2011), and the Japanese sentences are tokenized with the KyTea tokenizer (Neubig et al., 2011). The Zh-En dataset is obtained from the TsinghuaAligner website[1]. We treat their evaluation set as the training data and use the test set in Liu and Sun (2015).

The De-En, En-Fr, Zh-En datasets contain the distinction between sure and possible alignment links. The statistics of these datasets are shown in Table 1. We use the Ja-En development set to tune the hyper-parameters.

Baselines. We compare our models with:

- fast_align (Dyer et al., 2013): a popular statistical word aligner which is a simple, fast reparameterization of IBM Model 2.

- eflomal (Östling and Tiedemann, 2016): an efficient statistical word aligner using a Bayesian model with Markov Chain Monte Carlo (MCMC) inference.

- GIZA++ (Och and Ney, 2003; Gao and Vogel, 2008): an implementation of IBM models. Following previous work (Zenkel et al., 2020), we use five iterations each for Model 1, the HMM model, Model 3 and Model 4.

- SimAlign (Sabet et al., 2020): a BERT-based word aligner that is not fine-tuned on any parallel data. The authors propose three alignment extraction methods and we implement their IterMax model with default parameters.

- Zenkel et al. (2020) and Chen et al. (2020): two state-of-the-art neural word aligners based on MT models.

---

[1]http://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html

Implementation Details. Our main results are obtained by using the probability thresholding method on the contextualized embeddings in the 8-th layer of multilingual BERT-Base (mBERT; Devlin et al. (2019)) and we will discuss this choice in our ablation studies. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2e-5 and the batch size is set to 8. Following Peters et al. (2019), we set α to 1.5 for α-entmax. The threshold c is set to 0 for α-entmax and 0.001 for *softmax* and optimal transport. Unless otherwise stated, β is set to 0. We mainly evaluate the model performance using Alignment Error Rate (AER).

3.2 Main Results

We first train our model on each individual language pair, then investigate if it is possible to train multilingual word aligners.

Bilingual Model Performance. From Table 2, we can see that our *softmax* model can achieve consistent improvements over the baseline models, demonstrating the effectiveness of our proposed method. Surprisingly, directly extracting alignments from mBERT (the *w/o fine-tuning* setting) can already achieve better performance than the popular statistical word aligner GIZA++ on 4 out of 5 settings, especially in the Zh-En setting where the size of parallel data is small.

Multilingual Model Performance. We also randomly sample 200k parallel sentence pairs from each language pair (except for Zh-En where we take all of its 40k parallel sentences) and concatenate them together to train multilingual word aligners. As shown in Table 2, the multilingually trained word aligners can achieve further improvements and they consistently outperform our bilingual word aligners and all the baselines even though the size of training data for each individual language pair is smaller. The results demonstrate that we can indeed obtain a neural word aligner that has state-of-the-art and robust performance across different language pairs. We also test the performance of our consistency optimization objective in this setting. We can see that incorporating this objective (β=1) can significantly improve the model performance on Ro-En, while it also deteriorates the Ja-En and Zh-En performance by a non-negligible margin. We find that this is because the CO objective can significantly improve the alignment recall while sacrificing the precisions, and our Ro-En dataset

	Component	De-En	Fr-En	Ro-En	Ja-En	Zh-En	Speed
Prob.	*softmax*	**17.4**	**5.6**	**27.9**	**45.6**	**18.1**	**33.22**
	α-entmax	18.1	**5.6**	29.0	46.3	18.4	32.36
OT	Cosine	24.4	15.7	33.7	54.0	31.1	3.36
	Dot Product	25.4	17.1	34.1	54.2	30.9	3.82
	Euclidean	20.7	15.1	33.3	53.2	29.8	3.05

Table 3: Comparisons of probability thresholding (Prob.) and optimal transport (OT) for alignment extraction. We try both *softmax* and α-entmax for probability thresholding and different cost functions for optimal transport. We measure both the extraction speed (#sentences/seconds) and the alignment quality (AER) on five language pairs, namely German-English (De-En), French-English (Fr-En), Romanian-English (Ro-En), Japanese-English (Ja-En), and Chinese-English (Zh-En). The best scores are in **bold**.

tends to favor models with high recall and the Ja-En and Zh-En datasets have an opposite tendency.

Zero-Shot Performance. In this paragraph, we want to find out how our models perform on language pairs that it has never seen during training. To this end, for each language pair, we train our model with data of all the other language pairs and test its performance on the target language pair. Results in Table 2 demonstrate that training our models with parallel data on *other* language pairs can still improve the model performance on the target language pair. This is a very important result, as it indicates that our model can be used as a off-the-shelf tool for multilingual word alignment for any language supported by the underlying embeddings, *regardless of whether parallel data has been used for training or not*.

3.3 Ablation Studies

In this part, we compare the performance of different alignment extraction methods, pre-trained embedding models and training objectives.

Alignment Extraction Methods. We first compare the performance of our two proposed alignment extraction methods, namely the probability thresholding and optimal transport techniques. We use the representations of the 8-th layer of mBERT following Sabet et al. (2020).

As shown in Table 3, probability thresholding methods can consistently outperform optimal transport by a large margin on the five language pairs. In addition, probability thresholding methods are much faster than optimal transport. *softmax* is marginally better than α-entmax, yet one advantage of α-entmax is that we do not need to manually set

Model	Layer	De-En	Fr-En	Zh-En
	7	18.7	6.1	19.1
mBERT	8	**17.4**	**5.6**	***18.1***
	9	18.8	6.1	20.1
	4	21.1	6.8	**25.3**
XLM-15 (MLM)	5	**20.4**	**6.1**	26.1
	6	23.2	7.7	33.3
	4	16.4	4.9	**18.6**
XLM-15 (MLM+TLM)	5	***16.2***	***4.7***	23.7
	6	18.8	5.7	26.2
	7	20.5	8.5	30.8
XLM-100 (MLM)	8	**19.8**	**8.2**	**28.6**
	9	19.9	8.8	29.3
	5	24.4	10.3	33.2
XLM-R	6	**23.1**	**9.2**	30.7
	7	24.7	11.5	**28.1**

Table 4: Comparisons of different LMs in terms of AER. We extract alignments using *softmax* and take representations from different layers of LMs. The best scores for each individual model are in **bold** and the overall best scores are in ***italicized bold***.

the threshold. Therefore, we use both *softmax* and α-entmax to obtain the main results.

Pre-trained Embedding Models. In this paragraph, we investigate the performance of three different types of pre-trained embedding models, including mBERT, XLM (Lample and Conneau, 2019) and XLM-R (Conneau et al., 2020). For XLM, we have tried its three released models: 1) XLM-15 (MLM) pre-trained with MLM and supports 15 languages; 2) XLM-15 (MLM+TLM) pre-trained with both the MLM and TLM objectives and supports 15 languages; 3) XLM-100 (MLM) pre-trained with MLM and supports 100 languages. We use *softmax* to extract the alignments.

Because XLM-15 does not support Japanese or Romanian, we only report the performance on the three other language pairs in Table 4. We take representations from different layers and report the performance of the best three layers. We can see that while XLM-15 (MLM+TLM) can achieve the best performance on De-En and Fr-En, the best layer is not consistent across language pairs. On the other hand, the optimal configurations for mBERT are consistent across language pairs. In addition, considering mBERT supports many more languages than XLM-15 (MLM+TLM), we will use mBERT in the following sections.

Training Objectives. We also conduct ablation studies on each of our training objectives. We can see from Table 5 that the self-training objective can best improve the model performance. Also,

the translation language modeling and parallel sentence identification objectives can marginally benefit the model. The masked language modeling objective, on the other hand, cannot always improve the model and can sometimes even deteriorate the model performance, possibly because the TLM objective already provides the model with sufficient supervision signals.

3.4 Analysis

We conduct several analyses to better understand our models. Unless otherwise stated, we perform experiments on the *softmax* model using mBERT.

Incorporating Supervised Signals. We investigate if our models can benefit from supervised signals. If we have access to word-level gold labels for word alignment, we can simply utilize them in our self-training objectives. Specifically, we can set A_{ij} in Equation 4 to 1 if and only if they are aligned. In our experimental settings, we have gold labels for all the Zh-En sentences and 653 sentences from the Ja-En development set. Table 6 demonstrates that training our models with as few as 653 labeled sentences can dramatically improve the alignment quality, and combining labeled and unlabeled parallel data can further improve the model performance. This analysis demonstrate the generality of our models as they can also be applied in semi-supervised settings.

Growing Heuristics. As stated in Section 2.1, because our alignment extraction methods essentially take the intersection of forward and backward alignments, growing heuristics can also be applied in our settings. The main motivation of growing heuristics is to improve the recall of the resulting alignments. While effective in statistical word aligners, as shown in Table 7, the growing heuristics only improve our alignment extraction method on the vanilla mBERT model in the Ro-En setting while degrading the model performance on all the other language pairs. After fine-tuning, the growing heuristics can only hurt the model performance, possibly because the self-training objective encourages the forward and backward alignments to be symmetrical. Based on these results, we do not adopt the growing heuristics in our models.

Annotation Projection. Word alignment has been a useful tool in cross-lingual annotation projection (Yarowsky et al., 2001; Nicolai and Yarowsky, 2019). Therefore, it would be inter-

Model	Objective	De-En	Fr-En	Ro-En	Ja-En	Zh-En
	All	15.3	4.4	22.6	37.9	13.6
	All w/o MLM	15.3	4.4	22.8	38.6	13.7
softmax	*All w/o TLM*	15.5	4.7	22.9	39.7	14.0
	All w/o SO	16.9	4.8	23.0	39.1	15.4
	All w/o PSI	15.4	4.4	22.7	37.9	13.8

Table 5: Ablation studies on our training objectives in multilingual settings.
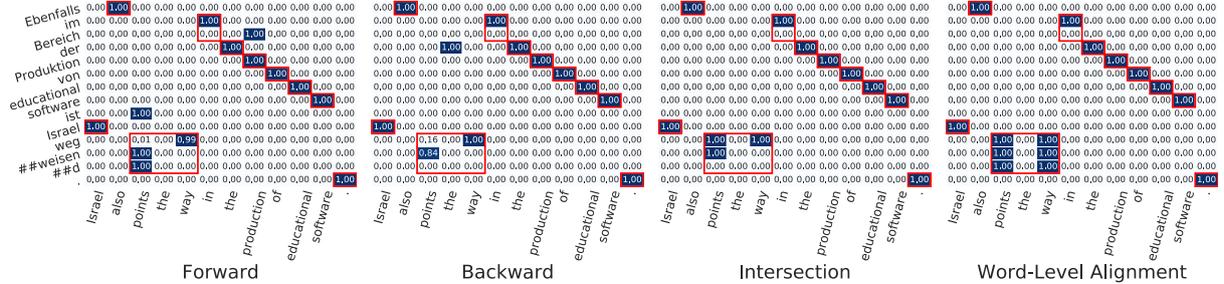
Figure 3: An example of extracting alignments from our fine-tuned model using *softmax*. Red boxes indicate the gold alignments. The fine-tuned model can generate more accurate alignments then vanilla mBERT (Figure 2).

Lang.	Unsup.	Sup.	Semi-Sup.
Zh-En	15.3	12.5	-
Ja-En	38.4	31.6	30.0

Table 6: Incorporating supervised word alignment signals into our model can further improve the model performance in terms of AER.

Model	Ext.	De-En	Fr-En	Ro-En	Ja-En	Zh-En
	X-En	24.7	14.4	31.9	54.7	27.4
	En-X	22.6	12.2	32.0	52.7	29.9
mBERT	*softmax*	17.4	5.6	27.9	45.6	18.1
	gd	18.7	9.2	27.0	48.5	23.4
	gd-final	18.6	9.3	26.9	48.7	23.2
	X-En	20.2	12.9	25.4	42.1	19.3
	En-X	18.1	9.3	25.9	41.7	23.5
Ours-Multi.	*softmax*	15.3	4.4	22.6	37.9	13.6
	gd	16.3	8.1	23.1	38.2	18.3
	gd-final	16.5	8.3	23.2	38.7	18.5

Table 7: The *grow-diag-final* heuristic can only improve our alignment extraction method in the Romanian-English setting without fine-tuning. "gd" refers to grow-diag.

Model	Prec. %	Rec. %	F_1 %
BERT-En (zero-shot)	53.1	54.3	52.7
fast_align	51.5	59.8	55.2
GIZA++	56.5	64.1	60.0
SimAlign	59.9	67.6	63.5
Ours	**60.6**	**68.5**	**64.3**

Table 8: Our model is also effective in an annotation projection setting where we train a BERT-based NER model on English data and test it on Spanish data. The best scores are in **bold**.

ing word aligners. From Table 8, we can see that our model is also better than baselines in this setting, demonstrating its usefulness in cross-lingual annotation projection.

Sentence-Level Representation Transfer. We also test if the aligned representations are beneficial for sentence-level cross-lingual transfer. In doing so, we perform experiments on XNLI (Conneau et al., 2018), which evaluates cross-lingual sentence representations in 15 languages on the task of natural language inference (NLI). We train our models with the provided 10k parallel data on the 15 languages, fine-tune our model on the English NLI data, then test its performance on other languages. As shown in Table 9, our model can outperform the baseline, indicating the aligned word representations can also be helpful for sentence-level cross-lingual transfer.

esting to see if our model can be beneficial in these settings. To this end, we evaluate our model and baselines on cross-lingual named entity recognition (NER). We train a BERT-based NER model on the CoNLL 2003 English data (Tjong Kim Sang and De Meulder, 2003) and test it on the CoNLL 2002 Spanish data (Tjong Kim Sang, 2002). We use Google Translate to translate Spanish test set into English, predict the labels using the NER model, then project the labels from English to Spanish us-

Model	En	Fr	Es	De	El	Bg	Ru	Tr	Ar	Vi	Th	Zh	Hi	Sw	Ur	Ave.
mBERT	81.3	73.4	74.3	70.5	66.9	68.2	68.5	59.5	64.3	**70.6**	50.7	68.8	59.3	49.4	57.5	65.5
Ours	**81.5**	**74.1***	**74.9***	**71.2***	**67.1**	**68.7***	**68.6**	**61.0***	**66.2***	70.5	**53.8***	**69.1**	**59.8***	**50.6***	**58.6***	**66.4***

Table 9: Results of mBERT and our fine-tuned model on XNLI (Conneau et al., 2018). Our objectives can improve the model cross-lingual transfer ability. "*" denotes significant differences using paired bootstrapping (p<0.05).

Alignment Examples. We also conduct qualitative analyses as shown in Figure 1, 2 and 3. After fine-tuning, the learned contextualized representations are more aligned, as the cosine distances between semantically similar words become closer, and the extracted alignments are more accurate. More examples are shown in Appendix B.

4 Related Work

Based on the IBM translation models (Brown et al., 1993), many statistical word aligners have been proposed (Vogel et al., 1996; Östling and Tiedemann, 2016), including the current most popular tools GIZA++ (Och and Ney, 2000, 2003; Gao and Vogel, 2008) and fast_align (Dyer et al., 2013).

Recently, there is a resurgence of interest in neural word alignment (Tamura et al., 2014; Alkhouli et al., 2018). Based on NMT models trained on parallel corpora, researchers have proposed several methods to extract alignments from them (Luong et al., 2015; Zenkel et al., 2019; Garg et al., 2019; Li et al., 2019) and successfully build an end-to-end neural model that can outperform statistical tools (Zenkel et al., 2020). However, there is an inherent discrepancy between translation and word alignment: translation models are directional and the source and target side are treated differently, while word alignment is a non-directional task. Therefore, certain adaptations are required for translation models to perform word alignment.

Another disadvantage of MT-based word aligners is that they cannot easily utilize contextualized embeddings. Using learned representations to improve word alignment have been investigated (Sabet et al., 2016; Pourdamghani et al., 2018). Recently, pre-trained LMs (Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020) have proven to be useful in cross-lingual transfer (Libovický et al., 2019; Hu et al., 2020b). In word alignment, Sabet et al. (2020) propose effective methods to extract alignments from multilingual LMs without explicit training on parallel data. In this work, we propose better alignment extraction methods and combine the best of the two worlds by fine-tuning contextualized embeddings on parallel data.

There are also work on supervised neural word alignment (Stengel-Eskin et al., 2019; Nagata et al., 2020). However, supervised data are not always accessible, making their methods inapplicable in many scenarios. In this paper, we demonstrate that our model can incorporate supervised signals if available and perform semi-supervised learning, which is a more realistic and general setting.

Some work on bilingual lexicon induction also share similar general ideas with ours. For example, Zhang et al. (2017) minimize the earth mover's distance to match the embedding distributions from different languages. Similarly, Grave et al. (2019) present an algorithm to align point clouds with Procrustes (Schönemann, 1966) in Wasserstein distance for unsupervised embedding alignment.

5 Discussion and Conclusion

We present a neural word aligner that achieves state-of-the-art performance on five diverse language pairs and obtains robust performance in zero-shot settings. We propose to fine-tune multilingual embeddings with objectives suitable for word alignment and develop two alignment extraction methods. We also demonstrate its applications in semi-supervised settings. We hope our word aligner can be a tool that can be used out-of-the-box with good performance over various language pairs. Future directions include designing better training objectives and experimenting on more language pairs.

Also, note that we mainly evaluate our word aligners using AER following previous work, which has certain limitations. For example, it may not be well-correlated with statistical machine translation performance Fraser and Marcu (2007) and different types of alignments can be suitable for different tasks or conditions (Lambert et al., 2012; Stymne et al., 2014). Although we have evaluated models in annotation projection and cross-lingual transfer settings, alternative metrics (Tiedemann, 2005; Søgaard and Wu, 2009; Ahrenberg, 2010) are also worth considering in the future.

Acknowledgement

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*.

Lars Ahrenberg. 2010. Alignment-based profiling of europarl data in an english-swedish parallel corpus. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Conference on Machine Translation*.

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

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Rrepresentations*.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Identifying and controlling important neurons in neural machine translation. In *Proceedings of the International Conference on Learning Representations*.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*.

Tom B. Brown, Benjamin Pickman Mann, Nick Ryder, Melanie Subbiah, Jean Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric J Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual alignment of contextual word representations. In *Proceedings of the International Conference on Learning Representations*.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Proceedings of the Advances in Neural Information Processing Systems*.

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 Conference of the North American Chapter of the 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 Conference of the North American Chapter of the Association for Computational Linguistics*.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *Proceedinds of the International Conference on Artificial Intelligence and Statistics*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Jonathan Herzig and Jonathan Berant. 2018. Decoupling structure and lexicon for zero-shot semantic parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2020a. Explicit alignment objectives for multilingual bidirectional encoders. *arXiv preprint*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the International Conference on Machine Learning*.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime G Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning*.

Patrik Lambert, Simon Petitrenaud, Yanjun Ma, and Andy Way. 2012. What types of word alignment improve statistical machine translation? *Machine Translation*.

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

Joël Legrand, Michael Auli, and Ronan Collobert. 2016. Neural network-based word alignment through score aggregation. In *Proceedings of the Conference on Machine Translation*.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual BERT? *arXiv preprint*.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *Proceedings of the International Conference on Computational Linguistics*.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of Workshop on Building and Using Parallel Texts*.

Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Masaaki Nagata, Chousa Katsuki, and Masaaki Nishino. 2020. A supervised word alignment

method based on cross-language span prediction using multilingual BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: System Demonstrations*.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Garrett Nicolai and David Yarowsky. 2019. Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*.

Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the Annual Meeting of the 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 Conference of the North American Chapter of the Association for Computational Linguistics*.

Nima Pourdamghani, Marjan Ghazvininejad, and Kevin Knight. 2018. Using word vectors to improve word alignments for low resource machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2020. A retrieve-and-rewrite initialization method for unsupervised machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Findings*.

Masoud Jalili Sabet, Heshaam Faili, and Gholamreza Haffari. 2016. Improving word alignment of rare words with word embeddings. In *Proceedings of the International Conference on Computational Linguistics*.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*.

Anders Søgaard and Dekai Wu. 2009. Empirical lower bounds on translation unit error rate for the full class of inversion transduction grammars. In *Proceedings of the International Conference on Parsing Technologies*.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2014. Estimating word alignment quality for smt reordering tasks. In *Proceedings of the Workshop on Statistical Machine Translation*.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2014. Recurrent neural networks for word alignment model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Jörg Tiedemann. 2005. Optimization of word alignment clues. *Natural Language Engineering*.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of the International Conference on Computational Linguistics*.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the Conference on Natural Language Learning*.

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 Conference on Natural Language Learning*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*.

David Vilar, Maja Popović, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments? In *Proceedings of the International Workshop on Spoken Language Translation*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the International Conference on Computational Linguistics*.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Hainan Xu, Shuoyang Ding, and Shinji Watanabe. 2019. Improving end-to-end speech recognition with pronunciation-assisted sub-word modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013a. A lightweight and high performance monolingual word aligner. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013b. Semi-Markov phrase-based monolingual alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the International Conference on Human Language Technology Research*.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint*.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. 2019. Regularizing neural machine translation by target-bidirectional agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Model	Prec. %	Rec.%	F$_1$ %
Baseline			
Yao et al. (2013a)	91.3	82.0	86.4
Yao et al. (2013b)	90.4	81.9	85.9
Sultan et al. (2014)	**93.5**	82.6	87.6
Ours			
mBERT	87.0	89.0	88.0
Ours-Multilingual	87.0	89.3	88.1
Ours-Supervised	87.2	**89.8**	**88.5**

Table 10: Our model is also effective in monolingual alignment settings.

A Implementation Details

We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2e-5 and the batch size is set to 8. Following Peters et al. (2019), we set α to 1.5 for α-entmax. The threshold c is set to 0 for α-entmax and 0.001 for *softmax* and optimal transport. We train our models on one 2080 Ti for one epoch and it takes 3 to 24 hours for the model to converge depending on the size of the dataset. We evaluate the model performance using Alignment Error Rate (AER).

B Analysis

In this section, we conduct more analyses of our models.

Monolingual Alignment. We also investigate how our models perform in monolingual alignment settings. Previous methods (MacCartney et al., 2008; Yao et al., 2013a,b; Sultan et al., 2014) typically exploit external resources such as WordNet to tackle the problem. As shown in Table 10, mBERT can outperform previous methods in terms of recall and F$_1$ without any fine-tuning. Our multilingually fine-tuned model can achieve better recall and slightly better F$_1$ score than the vanilla mBERT model, and fine-tuning our model with supervised signals can achieve further improvements.

Sensitivity Analysis. We also conduct a sensitivity analysis on the threshold c for our *softmax* alignment extraction method. As shown in Table 11, our method is relatively robust to this threshold. In particular, after fine-tuning, the AERs change within 0.5% when varying the threshold.

Comparisons with IterMax. IterMax is the best alignment extraction method in SimAlign (Sabet et al., 2020). The results in the main paper have demonstrated that our alignment extraction methods are able to outperform IterMax. In Figure 4, we

Model	c.	De-En	Fr-En	Ro-En	Ja-En	Zh-En
	1e-6	17.3	6.0	27.2	45.2	18.9
	1e-5	17.3	5.9	27.4	45.1	18.6
	1e-4	17.3	5.7	27.6	45.3	18.3
mBERT	1e-3	17.4	5.6	27.9	45.6	18.1
	1e-2	17.7	5.6	28.4	45.8	18.2
	1e-1	18.1	5.6	28.9	46.3	18.3
	5e-1	18.4	5.6	29.5	47.0	18.7
	1e-6	15.4	4.6	22.7	38.2	14.1
	1e-5	15.4	4.5	22.7	38.1	14.0
	1e-4	15.3	4.5	22.6	37.9	13.9
Ours-Multilingual	1e-3	15.3	4.4	22.6	37.9	13.8
	1e-2	15.3	4.3	22.7	37.9	13.8
	1e-1	15.4	4.3	22.8	38.0	13.8
	5e-1	15.4	4.2	23.0	38.2	13.9

Table 11: Our *softmax* alignment extraction method is relatively robust to the threshold c.

can see that the IterMax algorithm tends to sacrifice precision for a small improvements in recall, while our model can generate more accurate alignments.

Ablation Studies on Training Objectives. Table 12 presents more ablation studies on our training objectives. We can see that the self training objective is the most effective one, with the translation language modeling objective being the second and the parallel sentence identification objective being the third. The masked language modeling objective can sometimes hurt the model performance, possibly because of the translation language modeling objective.

Experiments on More Language Pairs. We also test our alignment extraction methods on other language pairs following the setting of Sabet et al. (2020) without fine-tuning as shown in Table 13.[2]

More Qualitative Examples. In addition to the examples provided in the main text, we also present some randomly sampled samples in Figure 5. We can clearly see that our model learns more aligned representations than the baseline model.

---

[2]Their English-Persian dataset is unavailable at the time of writing the paper.

Model	Objective	De-En	Fr-En	Ro-En	Ja-En	Zh-En
Ours-Bilingual						
	All	16.1	4.1	23.4	38.6	15.4
	All w/o MLM	15.6	4.2	23.3	38.8	15.1
α-entmax	*All w/o TLM*	16.4	4.3	23.7	40.1	15.3
	All w/o SO	17.8	4.7	23.9	39.4	16.3
	All w/o PSI	16.5	4.2	23.1	38.5	15.4
	All	15.6	4.4	23.0	38.4	15.3
	All w/o MLM	15.5	4.2	23.2	38.9	14.9
softmax	*All w/o TLM*	15.9	4.5	23.7	40.1	15.1
	All w/o SO	17.4	4.7	23.2	38.6	16.3
	All w/o PSI	15.6	4.3	23.1	38.8	15.4
Ours-Multilingual						
	All	15.4	4.1	22.9	37.4	13.9
	All w/o MLM	15.1	4.2	22.8	37.8	13.7
α-entmax	*All w/o TLM*	16.4	4.4	23.3	39.7	14.4
	All w/o SO	17.5	4.6	23.6	40.0	15.6
	All w/o PSI	15.5	3.9	23.0	38.2	14.1
	All	15.3	4.4	22.6	37.9	13.6
	All w/o MLM	15.3	4.4	22.8	38.6	13.7
softmax	*All w/o TLM*	15.5	4.7	22.9	39.7	14.0
	All w/o SO	16.9	4.8	23.0	39.1	15.4
	All w/o PSI	15.4	4.4	22.7	37.9	13.8

Table 12: Ablation studies on training objectives.

Model	En-Cs	En-Hi
GIZA++	18.2	51.8
SimAlign	13.4	40.2
Ours (*softmax*, c=1e-3)	**12.3**	41.2
Ours (*softmax*, c=1e-5)	12.7	39.5
Ours (*softmax*, c=1e-7)	13.3	**39.2**

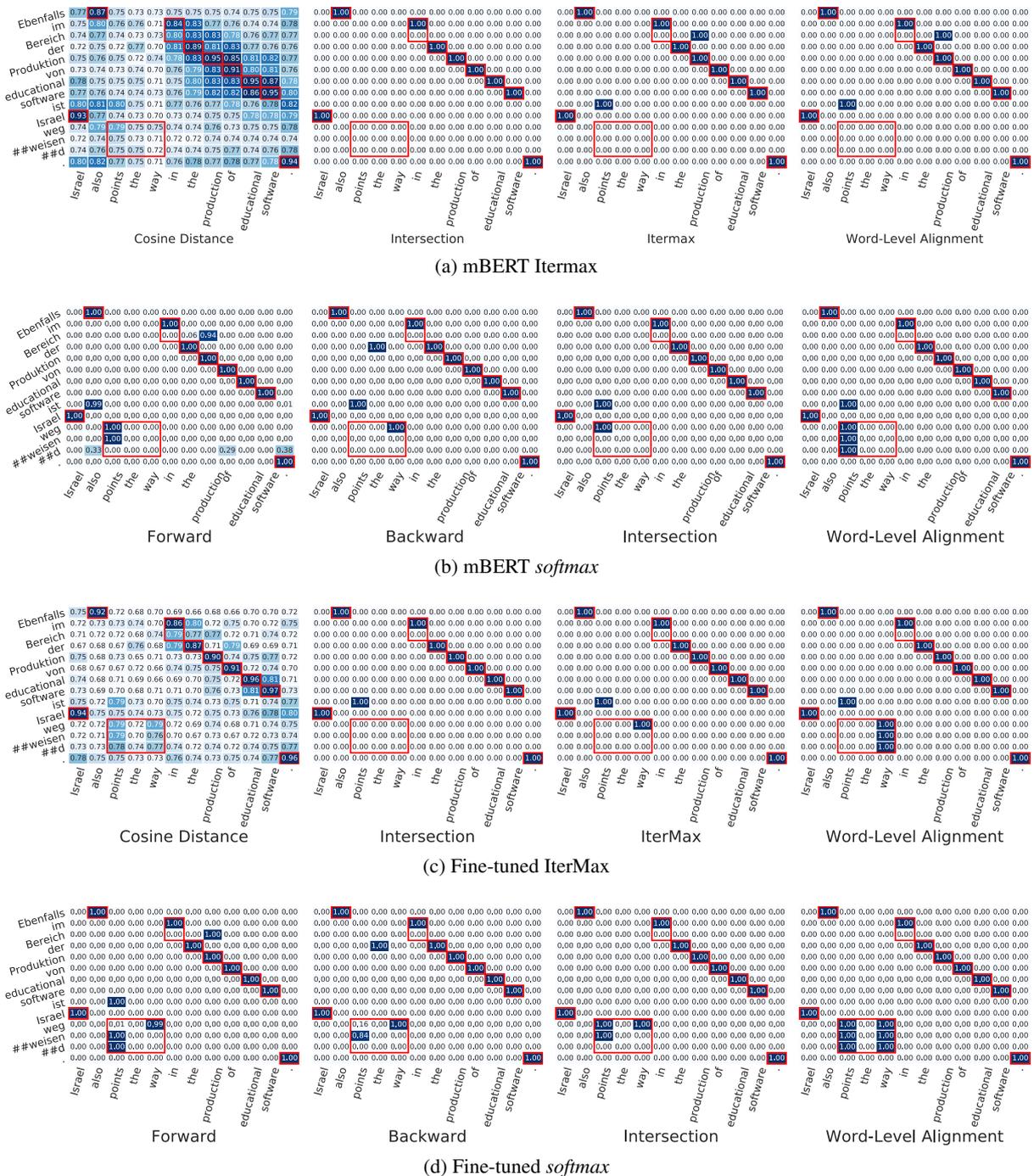Table 13: Performance on more language pairs.

Figure 4: Extracting alignments from our model using IterMax(Sabet et al., 2020) and our *softmax* method from the vanilla and fine-tuned mBERT models.
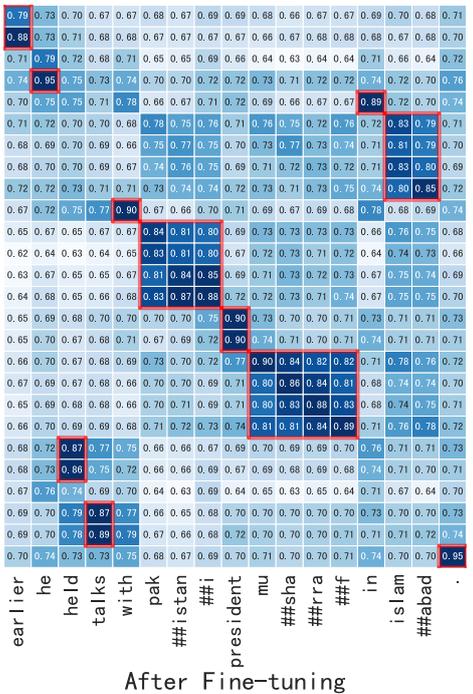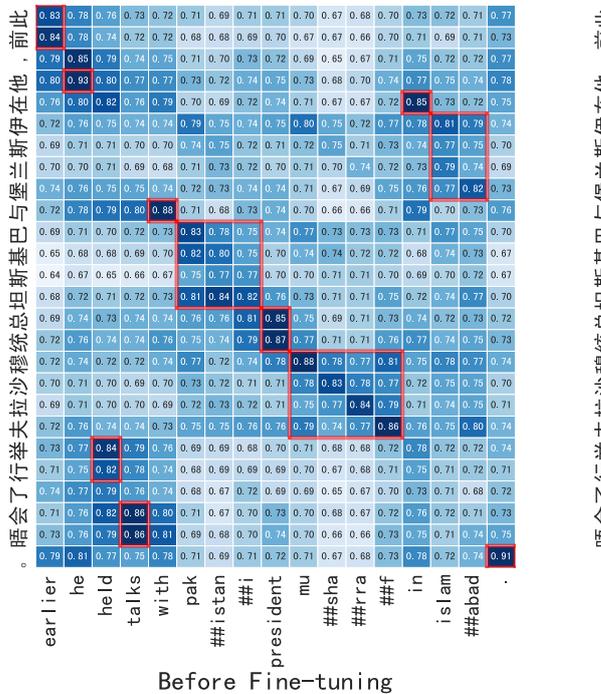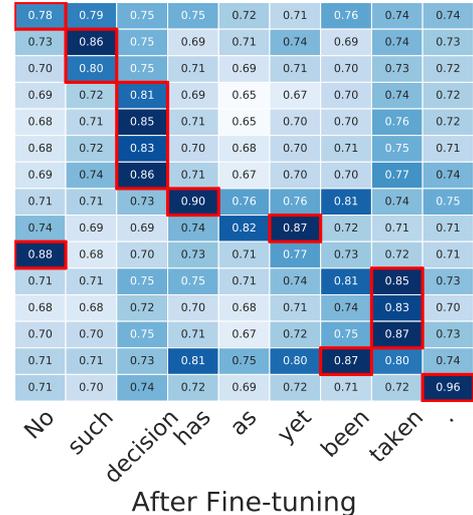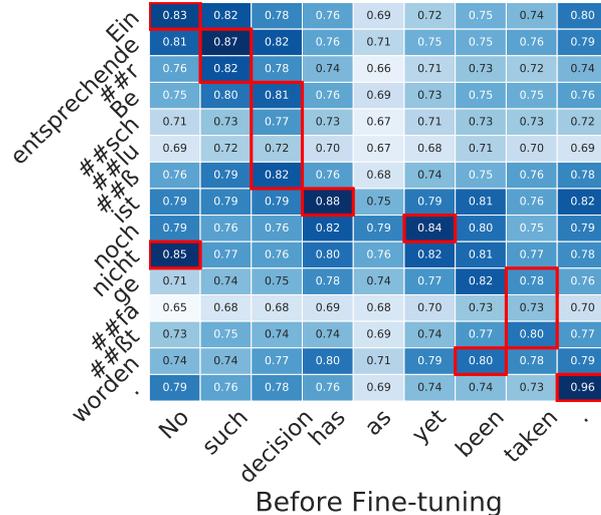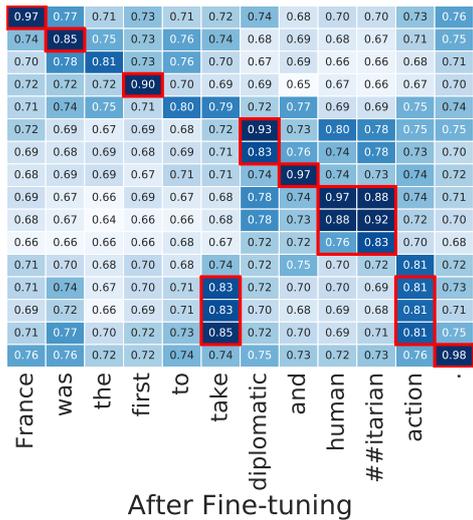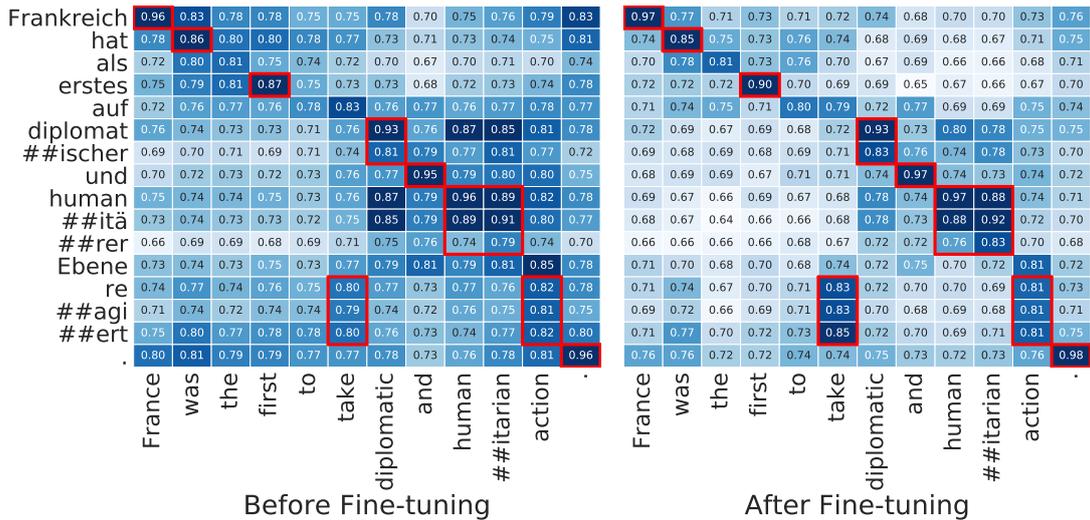
Figure 5: Cosine similarities between subword representations in a parallel sentence pair before and after fine-tuning. Red boxes indicate the gold alignments.