PMI-Align: Word Alignment With Point-Wise Mutual Information Without Requiring Parallel Training Data

Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti
School of Electrical and Computer Engineering, University of Tehran, Iran
{ft.azadi, hfaili, mjdousti}@ut.ac.ir

Abstract

Word alignment has many applications including cross-lingual annotation projection, bilingual lexicon extraction, and the evaluation or analysis of translation outputs. Recent studies show that using contextualized embeddings from pre-trained multilingual language models could give us high quality word alignments without the need of parallel training data. In this work, we propose PMI-Align which computes and uses the point-wise mutual information between source and target tokens to extract word alignments, instead of the cosine similarity or dot product which is mostly used in recent approaches. Our experiments show that our proposed PMI-Align approach could outperform the rival methods on five out of six language pairs. Although our approach requires no parallel training data, we show that this method could also benefit the approaches using parallel data to fine-tune pre-trained language models on word alignments. Our code and data are publicly available.

1 Introduction

Word alignment, as the task of finding the corresponding source and target tokens in a parallel sentence, was well-known as an essential component of statistical machine translation (SMT) systems. Despite the dominance of neural machine translation (NMT) in recent years, word alignment is still a notable area of research due to its usage in a wide variety of NLP applications, such as annotation projection (Yarowsky et al., 2001; Padó and Lapata, 2009; Huck et al., 2019; Nicolai and Yarowsky, 2019), bilingual lexicon extraction (Ammar et al., 2016; Shi et al., 2021; Artetxe et al., 2019), typological analysis (Lewis and Xia, 2008; Östling, 2015), guided alignment training of NMT (Liu et al., 2016; Chen et al., 2016; Alkhouli et al., 2018), and evaluation and analysis of translation outputs (Anthony et al., 2019; Neubig et al., 2019; Wang et al., 2020).

For many years statistical methods such as IBM models (Brown et al., 1993) and tools implemented based on them, namely GIZA++ (Och and Ney, 2003) or fast-align (Dyer et al., 2013), were among the most popular solutions to the word alignment task. Following the rise of deep neural models, several attempts have been made to extract word alignments from NMT models and their attention matrices (Peter et al., 2017; Ghader and Monz, 2017; Zenkel et al., 2020; Zhang and van Genabith, 2021). However, most of these methods, as well as the statistical aligners, require a sufficient amount of parallel training data to produce high quality word alignments. Recently, Jalili Sabet et al. (2020) have shown that high quality word alignments could be achieved using pre-trained multilingual language models (LMs), like MBERT Devlin et al. (2019) and XLMR Conneau et al. (2020). Their proposed method called SimAlign, extracts word alignments from similarity matrices induced from multilingual contextualized word embeddings with no need for parallel training data, which is very useful for low-resource language pairs. Afterwards, Dou and Neubig (2021) and Chi et al. (2021) proposed methods.
called probability thresholding and optimal transport to extract alignments using the similarity matrices derived from pre-trained LMs. They have also proposed some word alignment objectives to fine-tune the pre-trained models over parallel corpora.

In this paper, we follow the work done by Jalili Sabet et al. (2020) to extract alignments from pre-trained LMs without requiring any parallel training data and propose PMI-Align. Our main contribution is proposing to compute the \textit{point-wise mutual information} (PMI) between source and target tokens and using the PMI matrices instead of similarity matrices made of cosine similarities between the representation vectors of each source and target tokens, to align words. We argue that our proposed PMI-based method could align better as it considers the total alignment probability of each source or target token, as well as the joint alignment probabilities (equivalent to cosine similarities). This could alleviate the so-called hubness problem (Radovanovic et al., 2010) in high dimensional spaces, where some token’s representation is close to many others (see \_went in Figure 1). We perform experiments on six different language pairs and show that our method could surpass other alignment methods on five of them. We also conduct our experiments on different pre-trained LMs to show that PMI-Align could be advantageous regardless of the pre-trained model used.

2 Proposed Method

In this section, we first discuss how we define and compute the PMI matrix for each sentence pair and then we describe our alignment extraction method using the PMI matrix.

2.1 Point-Wise Mutual Information

Point-wise mutual information (PMI) is a well-known measure of association in information theory and NLP and it shows the probability of two events \( x \) and \( y \) occurring together, compared to what this probability would be if they were independent (Fano, 1961). It is computed as follows:

\[
\text{PMI}(x, y) := \log \frac{p(x, y)}{p(x)p(y)} \tag{1}
\]

In the context of word alignments, we define the PMI for a source and target token in a sentence pair as how more probable two tokens are to be aligned than if they are aligned randomly. Given a sentence \( x = < x_1, ..., x_n > \) in the source language and its corresponding target sentence \( y = < y_1, ..., y_m > \), the joint alignment probability of two tokens, \( x_i \) and \( y_j \), could be computed as:

\[
p(x_i, y_j) = \frac{e^{\text{sim}(h_{x_i}, h_{y_j})}}{\sum_{i', j'} e^{\text{sim}(h_{x_{i'}}, h_{y_{j'}})}}, \tag{2}
\]

where \( h_{x_i} \) is the contextualized embedding vector of \( x_i \) extracted from a pre-trained multilingual language model and \( \text{sim}(\cdot) \) is the cosine similarity measure. The total alignment probability of \( x_i \) and \( y_j \), i.e., \( p(x_i) \) and \( p(y_j) \), could also be computed according to the total probability rule as follows:

\[
p(x_i) = \sum_{1 \leq j' \leq m} p(x_i, y_{j'}) \tag{3}
\]

By calculating the PMI for each source and target token in a parallel sentence, we obtain the PMI matrix for that sentence pair, that could be used to extract alignments instead of similarity matrix in SimAlign (Jalili Sabet et al., 2020). The advantage of using PMI to align words is that it also considers the total alignment probability of each source and target token in addition to their joint alignment probability, which is equivalent to the similarity measure. This leads to reduce the probability to align the token pairs that one of them has high similarities to many other tokens, and thus could alleviate the so-called hubness problem in high dimensional spaces where some data points called hubs are the nearest neighbors of many others.

2.2 Extracting Alignments

To extract word alignments, we follow the simple Argmax method proposed in Jalili Sabet et al. (2020). Thus, we first obtain the source to target and target to source alignment matrices using the argmax over each row and each column of the PMI matrix, respectively. Next, we intersect these two matrices to get the final word alignment matrix. In other words, the final alignment matrix \( A_{ij} = 1 \) iff \( i = \text{argmax}_k(PMI_{ik}) \) and \( j = \text{argmax}_k(PMI_{jk}) \).

Since the above method would extract alignments on the subword level, we follow the heuristic used in previous work to obtain the word-level alignments by considering two words to be aligned if any of their subwords are aligned (Jalili Sabet et al., 2020; Zenkel et al., 2020; Dou and Neubig, 2021).
3 Experiments and Results

3.1 Datasets

We perform our experiments on six public datasets, as in (Jalili Sabet et al., 2020), consists of English-Czech (En-Cs), German-English (De-En), English-Persian (En-Fa), English-French (En-Fr), English-Hindi (En-Hi) and Romanian-English (Ro-En) language pairs. The statistics and URLs of these datasets are available in Table 2 in Appendix A.

3.2 Models and baselines

We compare our method with the following three state-of-the-art methods proposed to extract alignments from pre-trained multilingual LMs without using parallel training data. For all these methods default parameters were used in our experiments.

SimAlign\(^2\) (Jalili Sabet et al., 2020): They propose three methods to extract alignments from similarity matrices, called Argmax, Itermax and Match. Although Itermax and Match methods could not make significant improvements over Argmax and the Argmax method had better AER results for most of language pairs while using the XLMR-base model, they have argued that the Itermax method, which tries to apply Argmax iteratively, could be beneficial for more distant language pairs. Thus, we report both Argmax and Itermax results in our experiments to compare with our method.

Probability Thresholding\(^3\) (Dou and Neubig, 2021): In this method they apply a normalization function, i.e., softmax, to convert the similarity matrix of tokens into source to target and target to source alignment probability matrices. Afterwards, they extract the aligned words as the words that their alignment probabilities in both matrices exceed a particular threshold.

Optimal Transport\(^4\) (Chi et al., 2021): This method was proposed in both Dou and Neubig (2021) and Chi et al. (2021), and tried to model the word alignment task as the known optimal transport problem (Cuturi, 2013). Using the similarity matrix, this method attempted to find the alignment probability matrix that maximizes the sentence pair similarity. In our experiments, we use the method proposed by Chi et al. (2021) that utilizes the regularized variant of the optimal transport problem (Peyré et al., 2019), as it reported better results.

There are also many attempts made to improve the pre-trained LMs by fine-tuning on some parallel corpora to better align words. However, as our approach is irrelevant to the pre-trained model and our focus is on the alignment extraction instead of the model, we do not include those methods in our experiments. To demonstrate the effectiveness of our PMI-based alignment regardless of the utilized pre-trained multilingual LM, we conduct our experiments on M-BERT (Devlin et al., 2019), XLMR-Base (Conneau et al., 2020) and XLM-Align (Chi et al., 2021) which is fine-tuned on a word-alignment task, to show that our method could also be advantageous on more cross-lingually aligned models. All these models are publicly available in the Hugging Face platform (Wolf et al., 2020).

3.3 Results

Table 1 shows the results of our alignment technique compared to previous methods while using different pre-trained LMs. Following the previous work (Jalili Sabet et al., 2020; Dou and Neubig, 2021; Chi et al., 2021), we use the 8th layer’s representations of each pre-trained model to compute the similarity or PMI matrices. We also use the alignment error rate (AER) (Och and Ney, 2003) as the evaluation metric.

As Table 1 shows, our PMI-Align method could consistently outperform the other methods in all language pairs except En-Fr, regardless of the pre-trained model used. Compared to Argmax, our method performs better for about 1% or more in AER, while using the XLMR-Base model (except for En-Fr), which exclusively shows the benefits of using the PMI matrix instead of the similarity matrix. We also see that the PMI-Align could surpass the Itermax method for more distant language pairs such as En-Fa and En-Hi, where it was claimed to have the most advantage. Results show that our method could also be beneficial while using a model pre-trained on a word alignment task, i.e., XLM-align, which is expected to have more cross-lingually aligned representations, and less hubness problem.

The only language pair that our method could
### Table 1: AER results of our PMI-Align method compared to the other alignment extraction methods on 6 language pairs, while using different pre-trained models. The overall best results are in bold.

<table>
<thead>
<tr>
<th>Pretrained Model</th>
<th>Alignment method</th>
<th>En-Cs</th>
<th>De-En</th>
<th>En-Fa</th>
<th>En-Fr</th>
<th>En-Hi</th>
<th>Ro-En</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-BERT</td>
<td>SimAlign - Argmax</td>
<td>12.8</td>
<td>18.5</td>
<td>37.1</td>
<td>5.8</td>
<td>44.1</td>
<td>34.4</td>
<td>25.5</td>
</tr>
<tr>
<td></td>
<td>SimAlign - Itermax</td>
<td>15.0</td>
<td>19.0</td>
<td>33.8</td>
<td>9.0</td>
<td>41.3</td>
<td>31.2</td>
<td>24.9</td>
</tr>
<tr>
<td></td>
<td>Probability Thresholding</td>
<td>12.6</td>
<td>17.4</td>
<td>33.9</td>
<td>5.6</td>
<td>41.2</td>
<td>32.1</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>Optimal Transport</td>
<td>12.9</td>
<td>17.8</td>
<td>33.9</td>
<td>6.0</td>
<td>40.9</td>
<td>31.7</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>PMI-Align</td>
<td>11.8</td>
<td>17.0</td>
<td>32.8</td>
<td>5.7</td>
<td>39.3</td>
<td>30.9</td>
<td>20.9</td>
</tr>
<tr>
<td>XLM-Align</td>
<td>SimAlign - Argmax</td>
<td>10.7</td>
<td>16.6</td>
<td>28.4</td>
<td>5.6</td>
<td>34.6</td>
<td>27.7</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>SimAlign - Itermax</td>
<td>14.1</td>
<td>18.9</td>
<td>27.6</td>
<td>10.3</td>
<td>33.8</td>
<td>27.1</td>
<td>22.0</td>
</tr>
<tr>
<td></td>
<td>Probability Thresholding</td>
<td>13.7</td>
<td>18.5</td>
<td>29.6</td>
<td>7.9</td>
<td>35.2</td>
<td>28.4</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>Optimal Transport</td>
<td>11.1</td>
<td>16.6</td>
<td>28.0</td>
<td>6.6</td>
<td>34.0</td>
<td>27.0</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>PMI-Align</td>
<td>10.4</td>
<td>16.0</td>
<td>26.7</td>
<td>6.2</td>
<td>33.4</td>
<td>26.3</td>
<td>19.8</td>
</tr>
</tbody>
</table>

4 Related Work

Statistical aligners based on IBM models (Brown et al., 1993), such as Giza++ (Och and Ney, 2003) and fast align (Dyer et al., 2013) were the most dominant tools for word alignment until the late 2010s. With the rise of neural machine translation models, several attempts made to extract alignments from them (Ghader and Monz, 2017; Garg et al., 2019; Li et al., 2019; Zenkel et al., 2020; Chen et al., 2021; Zhang and van Genabith, 2021). However, all these models need parallel training data and could not utilize pre-trained contextualized embeddings. Recently, Jalili Sabet et al. (2020) have proposed methods to extract alignments from similarity matrices induced from multilingual LMs without the need for training on parallel data. Following this work, we propose a PMI measure to score and align words in each sentence pair, instead of cosine similarity. Some other alignment extraction methods using multilingual LMs were also provided by Dou and Neubig (2021) and Chi et al. (2021). They both also proposed several training objectives related to word alignments to fine-tune multilingual LMs on parallel data, as in some other recent works (Cao et al., 2020; Wu and Dredze, 2020; Lai et al., 2022).

5 Conclusions

This paper presents a word alignment extraction method based on the PMI matrices derived from cross-lingual contextualized embeddings, instead of just the similarity matrices. We proposed a way to compute the PMI matrix for each sentence pair and argued that using this PMI measure would be beneficial since for each source-target word pair, it considers not only their similarity to each other but also their similarity values to the other tokens of the sentence, that could mitigate the hubness problem.
Experimental results show that our PMI-Align method could outperform the previous alignment extraction methods in five out of six language pairs, regardless of the base pre-trained language model used to derive word embeddings. Although our method does not require any parallel training data, our experiments show that it could also benefit the approaches using such data to fine-tune the pre-trained models for better word alignments. In future work, the proposed PMI matrix could be investigated in other cross-lingual or even monolingual applications, like the translation quality estimation or the evaluation of text generation tasks, instead of the similarity matrix.

Limitations

Although our proposed aligner has surpassed the existing LM-based alignment extraction methods in most of the datasets, it could not make any improvement for the En-Fr language pair, as shown in Table 1. This suggests that our proposed method might be only beneficial for more distant languages. On the other hand, for similar languages, it not only cannot add any information to the similarity matrix, but also its estimation for the alignment probabilities might add noise to the alignment extraction method. Thus, investigating ways to more effectively estimate the alignment probabilities of source and target tokens might be helpful in future work.

Another limitation of our method, as well as other LM-based aligners, is that they first extract subword-level alignments, and then heuristically map them to word-level. By observing the aligner outputs, we realize that many errors occur when the pre-trained LM can not efficiently split words into meaningful subwords. This happens more often for low-resource languages or far languages from English (like Persian or Hindi). Thus, achieving better subword tokenization in pre-trained LMs or applicable methods to convert subword-level representations into word-level could help improve the quality of LM-based aligners.

References


Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In ICLR.


**A Data Statistics**

Table 2 shows the number of sentences and the download links of the test datasets we used in our experiments.

**B Alignment Examples**

Figures 2 and 3 illustrate some sentence pair examples comparing our PMI-Align method to SimAlign. They clearly show the advantages of using the PMI matrix over the similarity matrix. Both matrices are normalized with min-max normalization to be comparable.

**C Number of Parameters and Runtimes**

We use 3 pre-trained models in our experiments:

**MBERT** (Devlin et al., 2019), which is pre-trained with masked language modeling (MLM) and next sentence prediction on Wikipedia of 104 languages.

**XLMR-base** (Conneau et al., 2020), pre-trained with MLM on large-scale CommonCrawl data for 100 languages.

**XLM-align** (Chi et al., 2021), pre-trained with translation language modeling (TLM) and denoising word alignment (DWA) for 14 English-centric language pairs, along with MLM for 94 languages.

Our method has no parameters itself. However, considering the parameters of the used pre-trained LM, MBERT has about 170 million parameters, while XLMR-base and XLM-align both have about 270 million parameters.

Since our word aligner is simple and efficient, we did all our experiments on an Intel(R) Core(TM) i7-6700 CPU with 32GB memory and it just took about 0.1 seconds on average to align each parallel sentence in our whole dataset, while using XLMR-base model.

---

Table 2: Statistics and links for test datasets (Jalili Sabet et al., 2020)

<table>
<thead>
<tr>
<th>Language pair</th>
<th># of sentences</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-De</td>
<td>508</td>
<td><a href="http://www-i6.informatik.rwth-aachen.de/goldAlignment">http://www-i6.informatik.rwth-aachen.de/goldAlignment</a></td>
</tr>
<tr>
<td>En-Fr (Och and Ney, 2000)</td>
<td>447</td>
<td><a href="http://web.eecs.umich.edu/~mihalcea/wpt">http://web.eecs.umich.edu/~mihalcea/wpt</a></td>
</tr>
<tr>
<td>En-Hi</td>
<td>90</td>
<td><a href="http://web.eecs.umich.edu/~mihalcea/wpt05">http://web.eecs.umich.edu/~mihalcea/wpt05</a></td>
</tr>
<tr>
<td>En-Ro (Mihalcea and Pedersen, 2003)</td>
<td>203</td>
<td><a href="http://web.eecs.umich.edu/~mihalcea/wpt05">http://web.eecs.umich.edu/~mihalcea/wpt05</a></td>
</tr>
</tbody>
</table>
Figure 2: Similarity matrices (left) vs. PMI matrices (right) along with the word-level alignments extracted using SimAlign vs. PMI-Align for some parallel sentence pairs. Red boxes indicate the gold alignments, whereas red ovals show the aligners' outputs.
Figure 3: Additional examples.
The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Not applicable. Our method doesn’t have any hyperparameters.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Not applicable. Our results don’t vary in different runs.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

3.2

D  Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?

No response.

D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.