

# Aligning Very Small Parallel Corpora Using Cross-Lingual Word Embeddings and a Monogamy Objective

Nina Poerner, Masoud Jalili Sabet, Benjamin Roth and Hinrich Schütze

Center for Information and Language Processing

LMU Munich, Germany

poerner@cis.uni-muenchen.de

## Abstract

Count-based word alignment methods, such as the IBM models or fast-align, struggle on very small parallel corpora. We therefore present an alternative approach based on cross-lingual word embeddings (CLWEs), which are trained on purely monolingual data. Our main contribution is an unsupervised objective to adapt CLWEs to parallel corpora. In experiments on between 25 and 500 sentences, our method outperforms fast-align. We also show that our fine-tuning objective consistently improves a CLWE-only baseline.

## 1 Introduction

Some parallel corpora, such as the Universal Declaration of Human Rights, are too small to apply count-based word alignment algorithms.

Sabet et al. (2016) show that integrating monolingual word embeddings into IBM Model 1 (Brown et al., 1990) decreases word alignment error rate on a parallel corpus of 1000 sentences. Pourdamghani et al. (2018) exploit monolingual embedding similarity scores to create synthetic training data for Statistical Machine Translation (SMT), and report an increase in alignment F1.

Recent advances have made it possible to create cross-lingual word embeddings (CLWEs) from purely monolingual data (Zhang et al. (2017a), Zhang et al. (2017b), Conneau et al. (2017), Artetxe et al. (2018a)). We propose to leverage such CLWEs for a **similarity-based** word alignment method, which works on corpora as small as 25 sentences. Like Sabet et al. (2016), our method relies only on monolingual data (to train the embeddings) and on the small parallel corpus itself.

Our **CLWE-only baseline** aligns source and target words in a parallel corpus if their CLWEs have maximum cosine similarity. This approach is independent from the size of the parallel corpus, but has the following problems:

- Semantics may differ between the embedding training domain and the parallel corpus.
- CLWEs sometimes fail to discriminate between words with similar contexts, e.g., antonyms.

We therefore propose to **fine-tune** the CLWEs on the small parallel corpus using an **unsupervised embedding monogamy objective**. To evaluate the proposed method, we simulate sparse data settings using Europarl sentences and Bible verses. Our method outperforms the count-based fast-align model (Dyer et al., 2013) for corpus sizes up to 500 (resp., 250) sentences. The proposed fine-tuning method improves over the CLWE-only baseline in terms of both precision and recall.

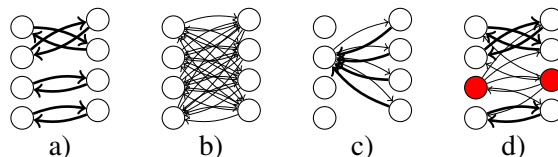


Figure 1: Schematic representation of the monogamy objective. a) one-to-one (“monogamous”) alignment:  $l(s, t) = 0$ , b) many-to-many alignment:  $l(s, t) = 1$ , c) one-to-many alignment:  $l(s, t) = 1$ , d) minimizing  $l(s, t)$  means making the red nodes more similar to each other, and less similar to the white nodes.

## 2 Method

### 2.1 CLWE-only baseline

Our CLWE-only baseline uses a cross-lingual embedding space derived from purely monolingual data (Artetxe et al., 2018a). Let  $D$  be our small corpus, and let  $s$  (source) and  $t$  (target) be parallel sentences from  $D$ . Let  $\text{clwe}(s_i)$  and  $\text{clwe}(t_j)$  be the embedding vectors of tokens  $s_i$  and  $t_j$ . We align  $s_i$  to  $\text{argmax}_{t_j \in t} [\cos(\text{clwe}(s_i), \text{clwe}(t_j))]$ .

Any ties are broken by proximity to the diagonal of the alignment matrix.

## 2.2 Fine-tuning method

**Intuition.** Assume that we have the following sentence pair: *aaa bbb xxx ||| 111 000 222*. Assume further that we know from CLWEs that *aaa*  $\approx$  *111* and *bbb*  $\approx$  *222*, but we lack informative embeddings for *000* and *xxx*. We may hypothesize that *xxx*  $\approx$  *000*, as they are the only tokens that lack translations. We may also hypothesize that *xxx*  $\not\approx$  *111*, *xxx*  $\not\approx$  *222*, as *111* and *222* already have translations of their own.

In the following, we will refer to this principle as **embedding monogamy**. We assume that in the absence of evidence to the contrary, a source embedding should have

- high similarity to one target embedding
- low similarity to other target embeddings<sup>1</sup>

This principle is related to the IBM Model (Brown et al., 1990), where Expectation Maximization increases  $p(f|e)$  if  $e$  and  $f$  co-occur in sentences where  $f$  is not explained by other source words.

**Embedding monogamy objective.** We define the probability of  $t_j$  given  $s_i$  as:

$$p(t_j|s_i, t) = \frac{e^{\frac{1}{\tau} \cos(\text{clwe}(s_i), \text{clwe}(t_j))}}{\sum_{j'} e^{\frac{1}{\tau} \cos(\text{clwe}(s_i), \text{clwe}(t_{j'}))}} \quad (1)$$

where  $\tau$  is a temperature hyperparameter. This definition is similar to the definition of translation probability in Artetxe et al. (2018b) and Lample et al. (2018). But while they normalize over the vocabulary, we normalize over the target sentence. As a consequence, the probability of  $t_j$  depends not only on  $s_i$ , but also on competitor tokens in  $t$ .

With these translation probabilities, we model a two-step random walker  $\mathbf{R}^{s \rightarrow t \rightarrow s}$  that starts at  $s_i$ , steps to a random target word and then to  $s_{i'}$ :  $r_{ii'}^{s \rightarrow t \rightarrow s} = \sum_{j=1}^{\text{len}(t)} p(t_j|s_i, t) p(s_{i'}|t_j, s)$ . To maximize monogamy, we maximize the entries on the diagonal of  $\mathbf{R}^{s \rightarrow t \rightarrow s}$ , i.e., the probability of the walker returning to its origin. To avoid penalizing long sentences, we minimize the negative logarithm to the base of the source sentence length:  $l(s, t) = 1 - \log_{\text{len}(s)} \sum_{i=1}^{\text{len}(s)} r_{ii}^{s \rightarrow t \rightarrow s}$ . This loss has the following properties:

<sup>1</sup> Of course, this assumption is over-simplistic, as one-to-n alignments exist (e.g., English *not* should be similar to both French *ne* and *pas*).

- In a fully “monogamous” situation (see Figure 1 a),  $r_{ii}^{s \rightarrow t \rightarrow s} \rightarrow 1 \implies l(s, t) \rightarrow 0$ .

- In a situation where all source words are equidistant from all target words (see Figure 1 b),  $r_{ii}^{s \rightarrow t \rightarrow s} = \frac{1}{\text{len}(s)} \implies l(s, t) = 1$ .

Reversing the roles of source and target results in the following bidirectional loss:  $L_{\text{bi}}(s, t) = \frac{1}{2}[l(s, t) + l(t, s)]$ . Both terms are necessary, since a given alignment may appear highly monogamous from the perspective of one sentence but not the other (especially when there are left-over words due to a difference in length).

**Adding position information.** At this point, our objective ignores word positions, which we know to be useful from count-based methods (e.g., Dyer et al. (2013)). Therefore, we add position embeddings inside the translation probability equation:

$$p(t_j|s_i, t) = \frac{e^{\frac{1}{\tau} \cos[\text{clwe}(s_i) + \mathbf{a}(i), \text{clwe}(t_j) + \mathbf{a}(j)]}}{\sum_{j'} e^{\frac{1}{\tau} \cos[\text{clwe}(s_i) + \mathbf{a}(i), \text{clwe}(t_{j'}) + \mathbf{a}(j')]}}$$

where  $\mathbf{a}(i)$  is a sinusoid embedding vector for position  $i$  (Vaswani et al., 2017). As a result, word pairs near the diagonal have higher round trip probabilities initially. Since the monogamy objective aims to strengthen strong links, similar position embeddings act as attractors for non-positional embeddings. Note that we use only the non-positional embeddings for alignment, as the position prior is too strong at test time.

**Alignment retention objective.** In initial experiments, we found that the monogamy objective increases recall but risks losing precision, relative to the CLWE-only baseline. Therefore, we add an additional objective that aims to increase round trip probability for alignments made by the baseline, but does not influence unaligned words:

$$L_{\text{ret}}(s, t) = \frac{1}{2}[l_{\text{ret}}(s, t) + l_{\text{ret}}(t, s)]$$

$$l_{\text{ret}}(s, t) = -\log \frac{\sum_{i,j} p(t_j|s_i, t) p(s_i|t_j, s) m_{ij}^{st}}{\sum_{i,j} m_{ij}^{st}}$$

$$m_{ij}^{st} = \mathbb{I}[(s_i, t_j) \in \text{align}_0]$$

where  $\text{align}_0$  is the intersection of the  $s$ -to- $t$  and  $t$ -to- $s$  alignments made with the initial CLWEs (see Section 2.1). Our final loss function is:  $L(D) = \frac{1}{|D|} \sum_{(s,t) \in D} [L_{\text{bi}}(s, t) + \alpha L_{\text{ret}}(s, t)]$ .

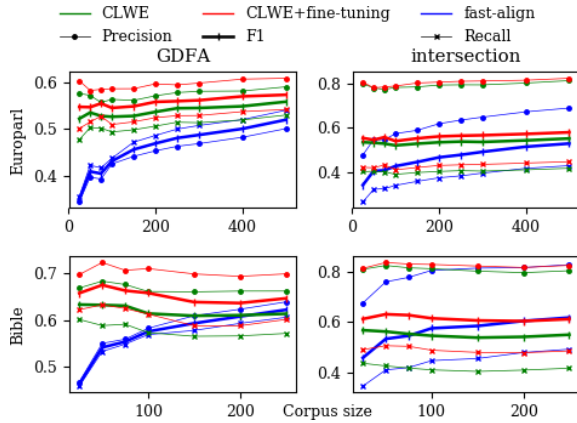


Figure 2: Alignment precision, recall and F1 as a function of corpus size.

### 3 Evaluation

We evaluate our model on subsets of different sizes from the English-German Europarl gold alignments<sup>2</sup> and French-English Bible gold alignments (Melamed, 1998)<sup>3</sup>. We initialize CLWEs with the unsupervised algorithm of Artetxe et al. (2018a) on monolingual FastText embeddings (Bojanowski et al., 2017)<sup>4</sup>. Fine-tuning is done in *keras*, using the adam optimizer (Kingma and Ba, 2014). We set  $\alpha = 1.0$  and  $\tau = 0.001$ , and apply 50% dropout to the embeddings.

We use fast-align (Dyer et al., 2013) as a count-based baseline, since it outperformed the IBM models in initial experiments. We symmetrize alignments by either intersection or the grow-diagonal-and (GDFA) heuristic (Koehn et al., 2007). We train fast-align and our fine-tuning method for 500 iterations.

## 4 Discussion

### 4.1 Corpus size

The performance of fast-align is highly dependent on corpus size, which is not surprising, seeing that it has to infer word semantics from the small corpus alone. The CLWE-only baseline on the other hand is independent from corpus size, resulting in decent performance even on 25 parallel sentences. Importantly, the positive effect of our fine-tuning method seems to be robust to corpus size, as we see improvements in F1 for all sizes.

<sup>2</sup>[www-i6.informatik.rwth-aachen.de/goldAlignment/](http://www-i6.informatik.rwth-aachen.de/goldAlignment/)

<sup>3</sup>[nlp.cs.nyu.edu/blinker/](http://nlp.cs.nyu.edu/blinker/). We consider links with inter-annotator agreement as sure, others as possible.

<sup>4</sup>[fasttext.cc](http://fasttext.cc), top-200000 words per language

### 4.2 Benefits of fine-tuning

We find that the proposed fine-tuning method has a positive effect on alignment precision and recall, relative to the CLWE-only baseline. We assess some sentence pairs qualitatively to find reasons for this improvement:

**Resolution of ambiguities.** Word embeddings sometimes fail to differentiate between words with similar contexts, such as antonyms. In Figure 3 (top), our fine-tuning method resolves such an ambiguity: Here, the initial CLWE of *answer* is slightly more similar to German *frage* (= *question*) than to the true translation *antwort*. Since *frage* already has a round trip partner, the monogamy objective pushes *answer* away from *frage*, resulting in the addition of a correct alignment between *answer* and *antwort*.

**In-domain word translations.** Since word embeddings are trained on general-purpose corpora, CLWEs can fail to reflect domain-specific word translations. One such example is the translation of *lord* as French *éternel* ( $\approx$  “*eternal one*”) in Figure 3 (bottom). While the translation is common in this particular Bible version, the CLWEs do not reflect it well ( $\cos(\text{lord}, \text{éternel}) < \cos(\text{wicked}, \text{éternel})$ ). Through fine-tuning, and due to their frequent cocurrence in the small corpus, the similarity between *éternel* and *lord* increases enough for a successful alignment.

## 5 Use case: Aligning the UDHR

In practice, our method would not be applied to English-German or English-French, as there is no lack of parallel data for these language pairs. For a more realistic use case, we align the 50 articles of the Universal Declaration of Human Rights<sup>5</sup> in Macedonian and Afrikaans. While we do not have gold alignments for an evaluation, a preliminary qualitative analysis suggests that our method finds a reasonable semantic word alignment, while fast-align mainly predicts the diagonal (see Figure 4 for examples).

## 6 Related Work

**Embeddings for word alignment.** Sabet et al. (2016) reformulate the IBM 1 model to predict the probability of monolingual target embedding vectors. They report improvements in AER for

<sup>5</sup><https://unicode.org/udhr/>

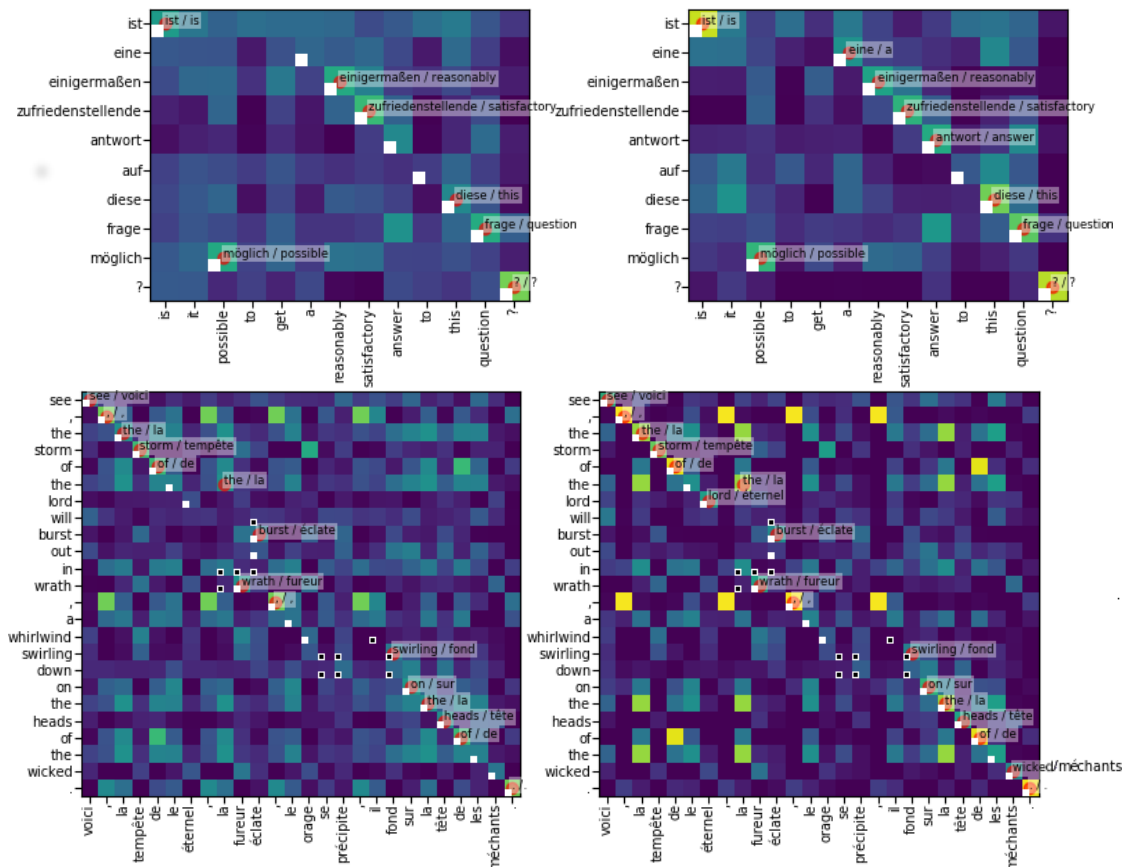


Figure 3: Similarity matrices before (left) and after (right) fine-tuning. Red dots: our alignment (intersection). White squares: sure gold alignments. Empty white squares: possible gold alignments.

English-French on parallel corpora between 1K and 40K sentences, as well as improvements in precision on words with frequency  $\leq 20$ .

Pourdamghani et al. (2018) exploit similarity scores from monolingual embeddings to create synthetic training data for an SMT system. They report improvements for English-Chinese, English-Arabic and English-Farsi alignment ( $\Delta F1 = 0.2\%, 0.5\%, 1.7\%$ ). Their smallest parallel corpus has 500K sentences, while we align a few dozen to hundred sentences.

**Two-step round trip objective.** Our use of two-step round trips is inspired by Haeusser et al. (2017). They optimize domain adaptation using a random walker that steps from image representations with known labels to image representations with unknown labels and back. While their target is a uniform distribution over images with the same label as the image of origin, ours is to have maximum probability mass on the word of origin.

**Low resource CLWEs.** Our approach relies on the availability of high-quality CLWEs. Wada and Iwata (2018) report that in settings with lit-

tle monolingual data ( $< 1M$  sentences), mapping approaches like Artetxe et al. (2018a) are not robust. Instead, they propose to learn CLWEs from a language model trained on the union of two small monolingual corpora. Their work is orthogonal to our fine-tuning method, since we make no assumptions about how the CLWEs are created.

## 7 Conclusion

We have presented a **similarity-based** method to produce word alignments for very small parallel corpora, using monolingual data and the corpus itself. Our **CLWE-only baseline** uses an unsupervised mapping of monolingual embeddings (Artetxe et al., 2018a). Our main contribution is an **unsupervised embedding monogamy objective**, which adapts CLWEs to the small parallel corpus. Our model outperforms count-based fast-align (Dyer et al., 2013) on parallel corpora up to 500 (resp., 250) sentences.

We expect that our method will be useful in low-resource settings, e.g., when aligning the Universal Declaration of Human Rights.

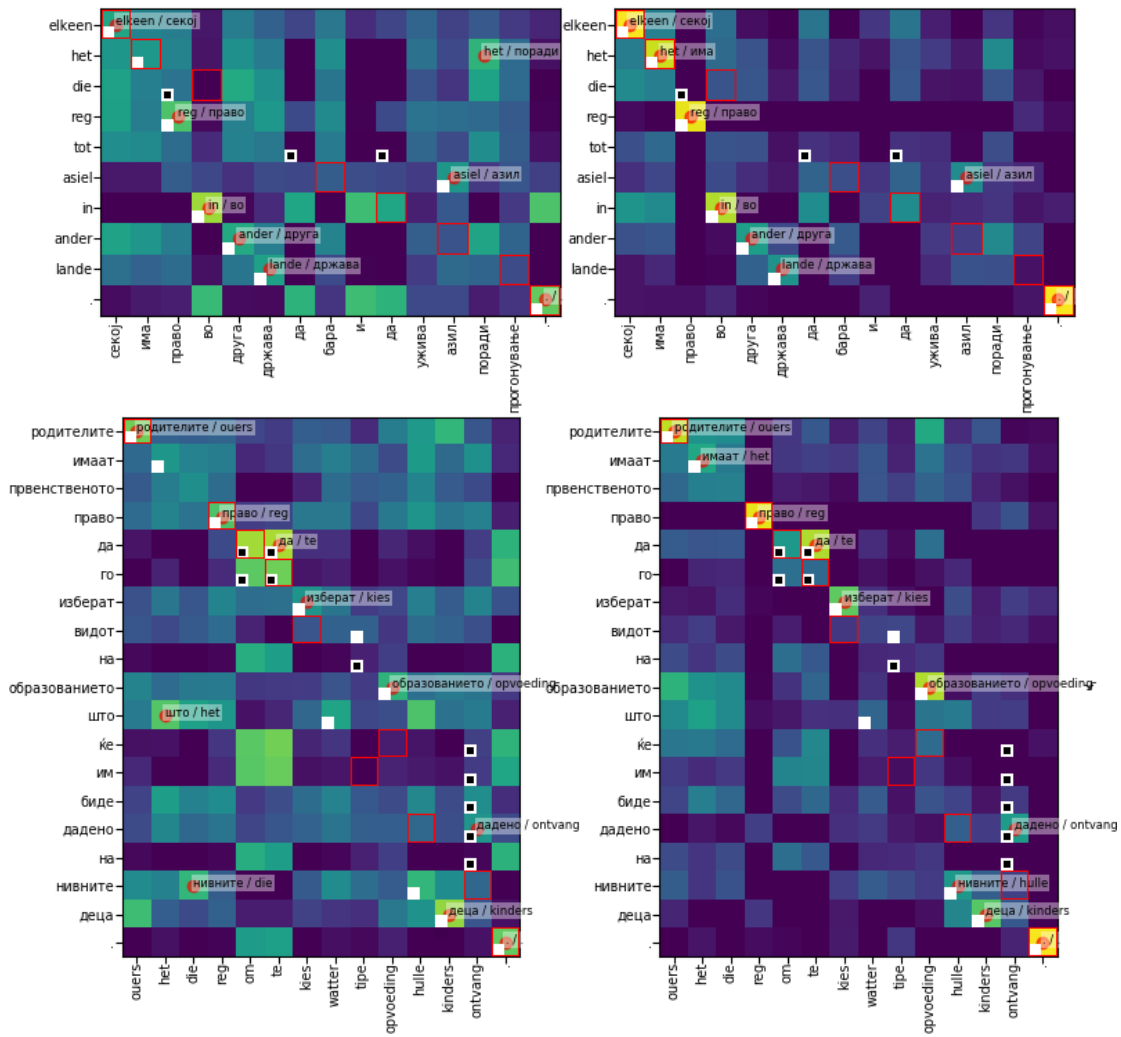


Figure 4: Articles 14(1) and 26(3) from the UDHR. Similarity matrices before (left) and after (right) fine-tuning. Red dots: our alignment (intersection). Red boxes: fast-align (intersection). White squares: sure gold alignments. Empty white squares: possible gold alignments (by the authors).

**Acknowledgments.** We gratefully acknowledge funding for this work by the European Research Council (ERC #740516).

## References

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, pages 789–798, Melbourne, Australia.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *EMNLP*, pages 3632–3642, Brussels, Belgium.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.
- Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *NAACL-HLT*, pages 644–648, Atlanta, USA.
- Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. 2017. Associative domain adaptation. In *ICCV*, pages 2765–2773, Venice, Italy.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180, Prague, Czech Republic.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.
- I Dan Melamed. 1998. Manual annotation of translational equivalence: The Blinker project. Technical report, University of Pennsylvania Institute for Research in Cognitive Science.
- Nima Pourdamghani, Marjan Ghazvininejad, and Kevin Knight. 2018. Using word vectors to improve word alignments for low resource machine translation. In *NAACL-HLT*, pages 524–528, New Orleans, USA.
- Masoud Jalili Sabet, Heshaam Faili, and Gholamreza Haffari. 2016. Improving word alignment of rare words with word embeddings. In *COLING 2016: Technical Papers*, pages 3209–3215, Osaka, Japan.
- 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 *NIPS*, pages 5998–6008, Long Beach, USA.
- Takashi Wada and Tomoharu Iwata. 2018. Unsupervised cross-lingual word embedding by multilingual neural language models. *arXiv preprint arXiv:1809.02306*.
- Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *ACL*, pages 1959–1970, Vancouver, Canada.
- Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In *EMNLP*, pages 1934–1945, Copenhagen, Denmark.