Unsupervised Word Alignment via Cross-Lingual Contrastive Learning

Anonymous authors
Paper under double-blind review

Abstract

Word alignment is essential for the down-streaming cross-lingual language understanding and generation tasks. Recently, the performance of the neural word alignment models (Zenkel et al., 2020b; Garg et al., 2019; Ding et al., 2019) has exceeded that of statistical models. However, they heavily rely on sophisticated translation models. In this study, we propose a super lightweight unsupervised word alignment model, dubbed MirrorAlign, in which a bidirectional symmetric attention trained with a contrastive learning objective is introduced, and an agreement loss is employed to bind the attention maps, such that the alignments follow mirror-like symmetry hypothesis. Experimental results on several public benchmarks demonstrate that our model achieves competitive, if not better, performance compared to the state of the art in word alignment while significantly reducing the training and decoding time on average. Further ablation analysis and case studies show the superiority of our proposed MirrorAlign. Notably, we recognize our model as a pioneer attempt to unify bilingual word embedding and word alignments. Encouragingly, our approach achieves 16.4× *speedup* against GIZA++, and 50× *parameter compression* compared with the Transformer-based alignment methods. We released our code to facilitate the community: `https://github.com/ICLR20anonymous/mirroralign`

1 Introduction

Word alignment, aiming to find the word-level correspondence between a pair of parallel sentences, is a core component of the statistical machine translation (SMT, Brown et al. 1993). It also has benefited several downstream tasks, *e.g.*, named-entity recognition (Che et al., 2013), part-of-speech tagging (Täckström et al., 2013), semantic role labeling (Kozhevnikov & Titov, 2013), cross-lingual dataset creation (Yarowsky et al., 2001) and cross-lingual modeling (Ding et al., 2020).

Recently, in the era of neural machine translation (NMT, Kalchbrenner & Blunsom 2013; Sutskever et al. 2014; Bahdanau et al. 2015; Gehring et al. 2017; Vaswani et al. 2017), the attention mechanism plays the role of the alignment model in translation system. Unfortunately, Koehn & Knowles (2017) show that attention mechanism may in fact dramatically diverge with word alignment. The works of (Li et al., 2019; Ghader & Monz, 2017) also confirm this finding.

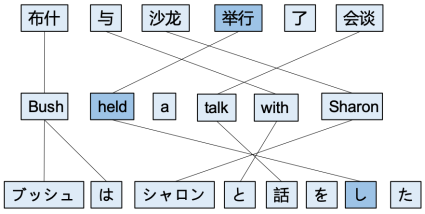


Figure 1: Two examples of word alignment. The upper and bottom cases are the Chinese and Japanese references, respectively.

Although there are some studies attempt to mitigate this problem, most of them are either rely on a sophisticated translation architecture (Zenkel et al., 2020b; Tang et al., 2019; Garg et al., 2019), or employed too much expensive human-annotated alignments (Stengel-Eskin et al., 2019b). As a result, statistical alignment tools, *e.g.,* FastAlign (Dyer et al., 2013) and GIZA++ (Och & Ney, 2003)[1], are still the most representative solutions due to its efficiency and unsupervised fashion. We argue that the word alignment task, intuitively, is much simpler than translation, and thus should be performed before

[1] GIZA++ employs the IBM Model 4 as default setting.

translation rather than inducing alignment matrix with heavy neural machine translation models. For example, The IBM word alignment model, *e.g.*, FastAlign, is the prerequisite of SMT. However, related research about lightweight neural word alignment without NMT is currently very scarce.

Inspired by cross-lingual word embeddings (CLWEs) (Luong et al., 2015), we propose to implement a lightweight unsupervised neural word alignment model, named MirrorAlign, which encourages the embeddings between aligned words to be closer. We also provide the theoretical justification from mutual information perspective for our proposed contrastive learning objective, demonstrating its reasonableness. As shown in Figure 1, if the Chinese word "举行" can be aligned to English word "held", the reverse mapping should also hold. Specifically, a bidirectional attention mechanism with contrastive learning objective is proposed to capture the alignment between parallel sentences. In addition, we employ an agreement loss to constrain the attention maps so that the alignments follow symmetry hypothesis (Liang et al., 2006).

Our contributions can be summarized as follows:

- We propose a bidirectional symmetric attention with contrastive learning objective for word alignment, in which we introduce extra loss function to follow the mirror-like symmetry hypothesis.
- We propose a lightweight unsupervised alignment structure, even merely updating the embedding matrices, achieves better alignment quality on several public benchmark datasets compare to baseline models while preserving comparable training efficiency with FastAlign.
- Further analysis show that the by-product of our model in training phase has the ability to learn bilingual word representations, which endows the possibility to unify these two tasks in the future.

2 Related Work

Word alignment studies can be divided into two classes:

Statistical Models Statistical alignment models directly build on the lexical translation models of Brown et al. (1993), also known as IBM models. The most popular implementation of this statistical alignment model is FastAlign (Dyer et al., 2013) and GIZA++ (Och & Ney, 2000; 2003). For optimal performance, the training pipeline of GIZA++ relies on multiple iterations of IBM Model 1, Model 3, Model 4 and the HMM alignment model (Vogel et al., 1996). Initialized with parameters from previous models, each subsequent model adds more assumptions about word alignments. Model 2 introduces non-uniform distortion, and Model 3 introduces fertility. Model 4 and the HMM alignment model introduce relative distortion, where the likelihood of the position of each alignment link is conditioned on the position of the previous alignment link. FastAlign (Dyer et al., 2013), which is based on a reparametrization of IBM Model 2, is almost the existing fastest word aligner, while keeping the quality of alignment.

In contrast to GIZA++, our MirrorAlign model achieves nearly 15× speedup during training, while achieving the comparable performance. Encouragingly, our model is at least 1.5× faster to train than FastAlign and consistently outperforms it.

Neural Models Most neural alignment approaches in the literature, such as Tamura et al. (2014) and Alkhouli et al. (2018), rely on alignments generated by statistical systems that are used as supervision for training the neural systems. These approaches tend to learn to copy the alignment errors from the supervising statistical models. Zenkel et al. (2019) use attention to extract alignments from a dedicated alignment layer of a neural model without using any output from a statistical aligner, but fail to match the quality of GIZA++. Garg et al. (2019) represents the current state of the art in word alignment, outperforming GIZA++ by training a single model that is able to both translate and align. This model is supervised with a guided alignment loss, and existing word alignments must be provided to the model during training. Garg et al. (2019) can produce alignments using an end-to-end neural training pipeline guided by attention activations, but this approach underperforms GIZA++. The performance of GIZA++ is only surpassed by training the guided alignment loss using GIZA++ output. Stengel-Eskin et al. (2019a) introduce a discriminative neural alignment model that uses a dot-product-based distance measure between learned source and target representation to predict if a given source-target pair should be aligned. Alignment decisions condition on the neighboring decisions using convolution. The model is trained using gold alignments. Zenkel et al. (2020a) uses

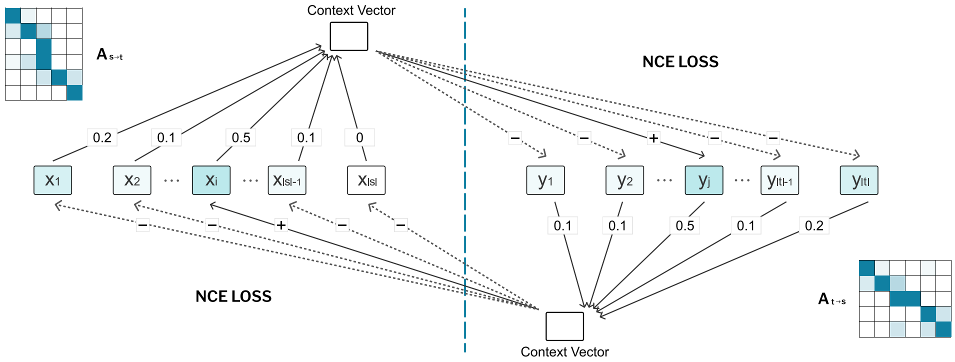


Figure 2: Illustration of the MirrorAlign, where a pair of sentences are given as example. Each x_i and y_j are the representation of words in source and target part respectively. Given y_j, we can calculate context vector in source part. The NCE training objective is encouraging the dot product of this context vector and y_j to be large. The process in the other direction is consistent. By stacking all of the soft weights, two attention maps $A_{s\to t}$ and $A_{t\to s}$ can be produced, which will be bound by an agreement loss to encourage symmetry.

guided alignment training, but with large number of modules and parameters, they can surpass the alignment quality of GIZA++.

They either use translation models for alignment task, which introduces a extremely huge number of parameters (compare to ours), making the training and deployment of the model cumbersome. Or they train the model with the alignment supervision, however, these alignment data is scarce in practice especially for low resource languages. These settings make above approaches less versatile.

Instead, our approach is fully unsupervised, that is, it does not require gold alignments generated by human annotators during training. Moreover, our model achieves comparable performance and is at least 50 times smaller than them, i.e., #Parameters: 4M (ours) vs. 200M (above).

3 Our Approach

Our model trains in an unsupervised fashion, where the word level alignments are not provided. Therefore, we need to leverage sentence-level supervision of the parallel corpus. To achieve this, we introduce negative sampling strategy with contrastive learning to fully exploit the corpus. Besides, inspired by the concept of cross-lingual word embedding, we design the model under the following assumption: *If a target token can be aligned to a source token, then the dot product of their embedding vectors should be large.* Figure-2 shows the schema of our approach **MirrorAlign**.

3.1 Sentence Representation

For a given source-target sentence pair $(\mathbf{s}, \mathbf{t})$, $s_i, t_j \in \mathbb{R}^d$ represent the i-th and j-th word embeddings for the source and target sentences, respectively. In order to capture the contextualized information of each word, we perform mean pooling operation with the representations of its surrounding words. Padding operation is used to ensure the sequence length. As a result, the final representation of each word can be calculated by element-wisely adding the mean pooling embedding and its original embedding:

$$x_i = \textsc{MeanPool}([s_i]^{win}) + s_i, \tag{1}$$

where win is the pooling window size. We can therefore derive the sentence level representations $(x_1, x_2, ..., x_{|s|}), (y_1, y_2, ..., y_{|t|})$ for $\mathbf{s}$ and $\mathbf{t}$.

3.2 Bidirectional Symmetric Attention

Bidirectional symmetric attention is the basic component of our proposed model. The aim of this module is to generate the source-to-target (*aka.* s2t) and target-to-source (*aka.* t2s) soft attention maps. The details of the attention mechanism: given a source side word representation x_i as query $q_i \in \mathbb{R}^d$ and pack all the target tokens together into a matrix $V_t \in \mathbb{R}^{|t| \times d}$. The attention context can be calculate as:

$$\textsc{Attention}\left(q_i, V_t, V_t\right) = (a_t^i \cdot V_t)^{\intercal}, \tag{2}$$

where the vector $a_t^i \in \mathbb{R}^{1 \times |t|}$ represents the attention probabilities for q_i in source sentence over all the target tokens, in which each element signifies the relevance to the query, and can be derived from:

$$a_t^i = \textsc{Softmax}\left(V_t \cdot q_i\right)^{\intercal}. \tag{3}$$

For simplicity, we denote the attention context of q_i in the target side as $att_t(q_i)$. s2t attention map $A_{s,t} \in \mathbb{R}^{|s| \times |t|}$ is constructed by stacking the probability vectors a_t^i corresponding to all the source tokens.

Reversely, we can obtain t2s attention map $A_{t,s}$ in a symmetric way. Then, these two attention matrices $A_{s,t}$ and $A_{t,s}$ will be used to decode alignment links. Take s2t for example, given a target token, the source token with the highest attention weight is viewed as the aligned word.

3.3 Agreement Mechanism

Intuitively, the two attention matrices $A_{s,t}$ and $A_{t,s}^T$ should be very close. However, the attention mechanism suffers from symmetry error in different direction (Koehn & Knowles, 2017). To bridge this discrepancy, we introduce agreement mechanism (Liang et al., 2006), acting like a mirror that precisely reflects the matching degree between $A_{s,t}$ and $A_{t,s}$, which is also empirically confirmed in machine translation (Levinboim et al., 2015; Cohn et al., 2016). In particular, we use an agreement loss to bind above two matrices:

$$\mathcal{L}oss_{disagree} = \sum_i \sum_j (A_{i,j}^{s,t} - A_{j,i}^{t,s})^2. \tag{4}$$

3.4 Training

Suppose that $(q_i, att_t(q_i))$ is a pair of s2t word representation and corresponding attention context sampled from the joint distribution $p_t(q, att_t(q))$ (hereinafter we call it a positive pair), the primary objective of the s2t training is to maximize the alignment degree between the elements within a positive pair. Thus, we first define an alignment function by using the sigmoid inner product as:

$$\textsc{Align}(q, att_t(q)) = \sigma(\langle q, att_t(q) \rangle), \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\langle \cdot, \cdot \rangle$ is the inner product operation. However, merely optimizing the alignment of positive pairs ignores important positive-negative relation knowledge (Mikolov et al., 2013).

To make the training process more informative, we reform the overall objective in the contrastive learning manner (Saunshi et al., 2019; Oord et al., 2018; Gutmann & Hyvärinen, 2010) with Noise Contrastive Estimation (NCE) loss (Mnih & Teh, 2012; Gutmann & Hyvärinen, 2012; Mikolov et al., 2013). Specifically, we first sample k negative word representations ${q_j}^2$ from the margin $p_t(q)$. Then, we can formulate the overall NCE objective as following:

$$\mathcal{L}oss_{s \to t}^i = - \mathop{\mathbb{E}}_{\{att_t(q_i), q_i, q_j\}} [\log \frac{\textsc{Align}(q_i, att_t(q_i))}{\textsc{Align}(q_i, att_t(q_i)) + \sum_{j=1}^k \textsc{Align}(q_j, att_t(q_i))}], \tag{6}$$

It is evident that the objective in Eq. (6) explicitly encourages the alignment of positive pair $(q_i, att_t(q_i))$ while simultaneously separates the negative pairs $(q_j, att_t(q_i))$. Moreover, a direct

[2] In the contrastive learning setting, q_j and $att_t(q_i)$ can be sampled from different sentences. If q_j and $att_t(q_i)$ are from the same sentence, $i \neq j$; otherwise, j can be a random index within the sentence length. For simplicity, in this paper, we use q_j where $i \neq j$ to denote the negative samples, although with a little bit ambiguity.

Method	EN-FR	FR-EN	sym	RO-EN	EN-RO	sym	DE-EN	EN-DE	sym
NNSA	22.2	24.2	15.7	47.0	45.5	40.3	36.9	36.3	29.5
FastAlign	16.4	15.9	10.5	33.8	35.5	32.1	28.4	32.0	27.0
MirrorAlign	**15.3**	**15.6**	**9.2**	34.3	**35.2**	**31.6**	31.1	**28.0**	**24.8**

Table 2: AER of each method in different direction. “sym” means grow-diag symmetrization.

consequence of minimizing Eq. (6) is that the optimal estimation of the alignment between the representation and attention context is proportional to the ratio of joint distribution and the product of margins $\frac{p_t(q,att_t(q))}{p_t(q)\cdot p_t(att_t(q))}$ which is the point-wise mutual information, and we can further have the following proposition with repect to the mutual information:

Proposition 1. *The mutual information between the word representation q and its corresponding attention context $att_t(q)$ is lower-bounded by the negative $\mathcal{L}oss^i_{s\to t}$ in Eq. (6) as:*

$$I(q, att_t(q)) \geq \log(k) - \mathcal{L}oss^i_{s\to t}, \tag{7}$$

where k is the number of the negative samples.

The detailed proof can be found in Oord et al. (2018) and Tian et al. (2019). Proposition 1 indicates that the lower bound of the mutual information $I(q, att_t(q))$ can be maximized by achieving the optimal NCE loss, which provides theoretical guarantee for our proposed method.

Our training schema over parallel sentences is mainly inspired by the bilingual skip-gram model (Luong et al., 2015) and invertibility modeling (Levinboim et al., 2015; Cohn et al., 2016). Therefore, the ultimate training objective should consider both forward ($s \to t$) and backward ($t \to s$) direction, combined with the mirror agreement loss. Technically, the final training objective is:

$$\mathcal{L}oss = \sum_{i}^{|t|} \mathcal{L}oss^i_{s\to t} + \sum_{j}^{|s|} \mathcal{L}oss^j_{t\to s} + \alpha \cdot \mathcal{L}oss_{disagree}, \tag{8}$$

where $\mathcal{L}oss_{s\to t}$ and $\mathcal{L}oss_{t\to s}$ are symmetrical and α is a loss weight to balance the likelihood and disagreement loss.

4 Experiments

4.1 Datasets and Evaluation Metrics

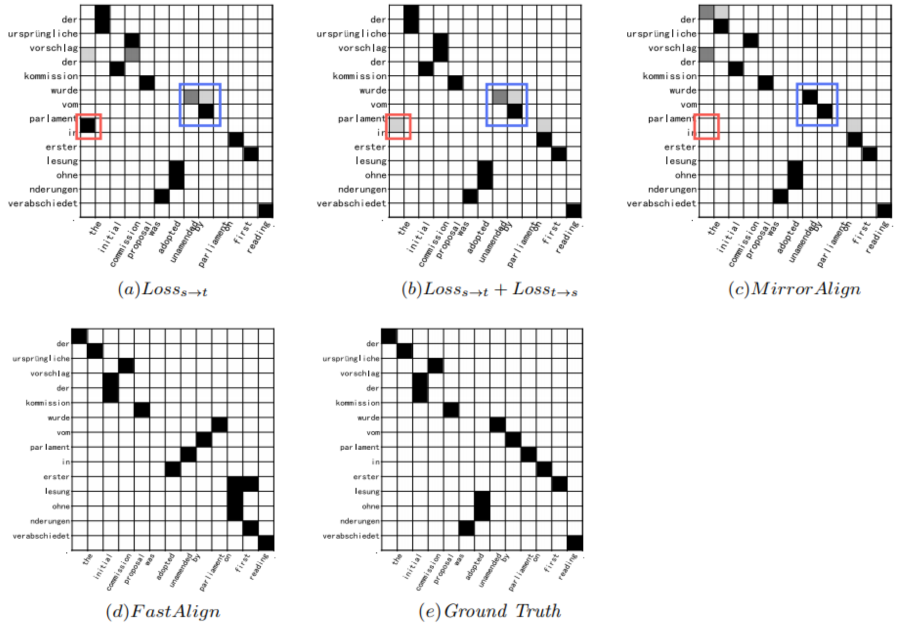
We perform our method on three widely used datasets: English-French (**EN-FR**), Romanian-English (**RO-EN**) and German-English (**DE-EN**). Training and test data for **EN-FR** and **RO-EN** are from NAACL 2003 share tasks (Mihalcea & Pedersen, 2003). For **RO-EN**, we merge Europarl v8 corpus, increasing the amount of training data from 49K to 0.4M. For **DE-EN**, we use the Europarl v7 corpus as training data and test on the gold alignments (Vilar et al., 2006). All above data are lowercased and tokenized by Moses. The evaluation metrics are Precision, Recall, F-score (F1) and Alignment Error Rate (AER) (Och & Ney, 2000).

Model	EN-FR	RO-EN	DE-EN
Naive Attention	31.4	39.8	50.9
NNSA	15.7	40.3	-
FastAlign	10.5	32.1	27.0
MirrorAlign	**9.2**	**31.6**	**24.8**
Zenkel et al. (2020b)	8.4	24.1	17.9
Garg et al. (2019)	7.7	26.0	20.2
GIZA++	5.5	26.5	18.7

Table 1: Alignment performance (with grow-diagonal heuristic) of each model.

4.2 Baseline Methods

FastAlign One of the most popular statistical method which log-linearly reparameterize the IBM model 2 proposed by Dyer et al. (2013).



$(a) Loss_{s \to t}$ $(b) Loss_{s \to t} + Loss_{t \to s}$ $(c) MirrorAlign$

$(d) FastAlign$ $(e) Ground\ Truth$

Figure 3: An visualized alignment example. (a-c) illustrate the effects when gradually adding the symmetric component, (d) shows the result of FastAlign, and (e) is the ground truth. The more emphasis is placed on the symmetry of the model, the better the alignment results model achieved. Meanwhile, as depicted, the results of the attention map become more and more diagonally concentrated.

GIZA++ A statistical generative model (Och & Ney, 2003), in which parameters are estimated using the Expectation-Maximization (EM) algorithm, allowing it to automatically extract bilingual lexicon from parallel corpus without any annotated data.

NNSA A unsupervised neural alignment model proposed by Legrand et al. (2016), which applies an aggregation operation borrowed from the computer vision to design sentence-level matching loss. In addition to the raw word indices, following three extra features are introduced: distance to the diagonal, part-of-speech and unigram character position. To make a fair comparison, we report the result of raw feature in NNSA.

Naive Attention Averaging all attention matrices in the Transformer architecture, and selecting the source unit with the maximal attention value for each target unit as alignments. We borrow the results reported in (Zenkel et al., 2019) to highlight the weakness of such naive version, where significant improvement are achieved after introducing an extra alignment layer.

Others (Garg et al., 2019) and (Zenkel et al., 2020b) represent the current developments in word alignment, which both outperform GIZA++. However, They both implement the alignment model based on a sophisticated translation model. Further more, the former uses the output of GIZA++ as supervision, and the latter introduces a pre-trained state-of-the-art neural translation model. It is unfair to compare our results directly with them. We still report their results in Table 1 as baselines.

4.3 SETUP

For our method (MirrorAlign), all the source and target embeddings are initialized by xavier method (Glorot & Bengio, 2010). The embedding size d and pooling window size are set to 256 and 3, respectively. The hyper-parameters α is tested by grid search from 0.0 to 1.0 at 0.1

intervals. For FastAlign, we train it from scratch by the open-source pipeline[3]. Also, we report the results of NNSA and machine translation based model(Sec.§4.2). All experiments of MirrorAlign are run on 1 `Nvidia K80` GPU. The CPU model is Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz. Both FastAlign and MirrorAlign take nearly half a hour to train one million samples.

4.4 Main Results

Table 1 summarizes the AER of our method over several language pairs. Our model outperforms all other baseline models. Comparing to FastAlign, we achieve 1.3, 0.5 and 2.2 AER improvements on **EN-FR**, **RO-EN**, **DE-EN** respectively.

Notably, our model exceeds the naive attention model in a big margin in terms of AER (ranging from 8.2 to 26.1) over all language pairs. We attribute the poor performance of the straightforward attention model (translation model) to its contextualized word representation. For instance, when translating a verb, contextual information will be paid attention to determine the form (*e.g.,* tense) of the word, that may interfere the word alignment.

Experiment results in different alignment directions can be found in Table 2. The grow-diag symmetrization benifits all the models.

4.5 Speed Comparison

Take the experiment on EN-FR dataset as an example, MirrorAlign converges to the best performance after running 3 epochs and taking 14 minutes totally, where FastAlign and GIZA++ cost 21 and 230 minutes, respectively, to achieve the best results. Notably, the time consumption will rise dozens of times in neural translation fashion. All experiments of MirrorAlign are run on a single Nvidia P40 GPU.

4.6 Ablation Study

To further explore the effects of several components (*i.e.,* bidirectional symmetric attention, agreement loss) in our MirrorAlign, we conduct an ablation study. Table 3 shows the results on **EN-FR** dataset. When the model is trained using only $\mathcal{L}oss_{s\rightarrow t}$ or $\mathcal{L}oss_{t\rightarrow s}$ as loss functions, the AER of them are quite high (20.9 and 23.3). As expected, combined loss function improves the alignment quality significantly (14.1 AER). It is noteworthy that with the rectification of agreement mechanism, the final combination achieves the best result (9.2 AER), indicating that the agreement mechanism is the most important component in MirrorAlign.

To better present the improvements brought by adding each component, we visualize the alignment case in Figure-3. The attention map becomes more diagonally concentrated after adding the bidirectional symmetric attention and the agreement constraint.

5 Analysis

Alignment Case Study We analyze an alignment example in Figure- 4. Compared to FastAlign, our model correctly aligns "*do not believe*" in English to "*glauben nicht*" in German. Our model, based on word representation, makes better use of semantics to accomplish alignment such that inverted phrase like "*glauben nicht*" can be well handled. Instead, FastAlign, relied on the positional assumption[4], fails here.

Setup	P	R	F1	AER
$\mathcal{L}oss_{s\rightarrow t}$	74.9	86.0	80.4	20.9
$\mathcal{L}oss_{t\rightarrow s}$	71.9	85.3	77.3	23.3
$\mathcal{L}oss_{s\leftrightarrow t}$	81.5	**90.1**	86.1	14.1
MirrorAlign	**91.8**	89.1	**90.8**	**9.2**

Table 3: Ablation results on EN-FR dataset.

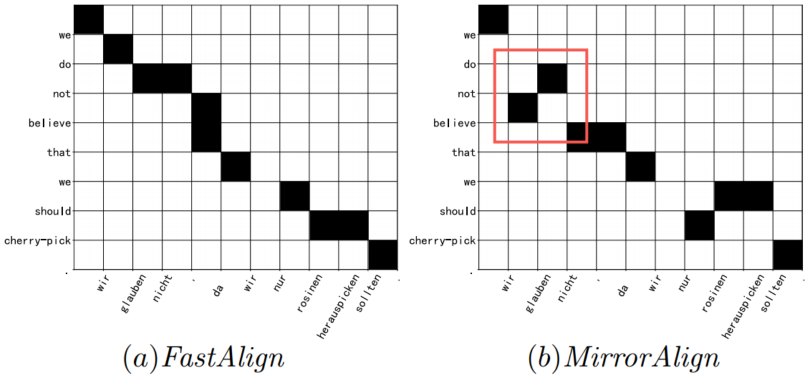
Word Embedding Clustering To further investigate the effectiveness of our model, we also analyze the word embeddings learned by our model. In

[3] https://github.com/lilt/alignment-scripts

[4] A feature h of position is introduced in FastAlign to encourage alignments to occur around the diagonal. $h(i,j,m,n) = -\left|\frac{i}{m} - \frac{j}{n}\right|$, i and j are source and target indices and m and n are the length of sentences pair.

china		distinctive		easily	
EN	**DE**	**EN**	**DE**	**EN**	**DE**
china	chinas	distinctive	unverwechselbaren	easily	unschwer
chinese	china	distinct	besonderheiten	easy	m‘helos
china’s	chinesische	peculiar	markante	readily	leichtes
republic	chinesischer	differences	charakteristische	starightforward	einfacheren
india	porzellanladen	influential	besonderheit	lightly	leicht
cat		love		january	
EN	**DE**	**EN**	**DE**	**EN**	**DE**
cat	hundefelle	love	liebe	january	j‘nner
dog	katzenfell	affection	liebt	october	januar
toys	hundefellen	loved	liebe	march	januartagen
grandchildren	katze	fond	geliebt	june	1.1.2002
cats	k‘chen	loves	lieben	july	15.januar

Table 4: Nearest neighbor words - shown are the top 10 nearest English (**EN**) and German (**DE**) words for each of the following words: *china*, *distinctive*, *easily*, *cat*, *love*, *january*.



$(a) FastAlign$ $(b) MirrorAlign$

Figure 4: Example of the DE-EN alignment. (a) is the result of FastAlign, and (b) shows result of our model, which is closer to the gold alignment. The horizontal axis shows German sentence “*wir glauben nicht , da wir nur rosinen herauspicken sollten .*”, and the vertical axis shows English sentence “*we do not believe that we should cherry-pick .*”.

particular, following Collobert et al. (2011), we show some words together with its nearest neighbors using the Euclidean distance between their embeddings. Table- 4 shows some examples to demonstrates that our learned representations possess a clearly clustering structure bilingually and monolingually.

We attribute the better alignment results to the ability of our model that could learn bilingual word representation.

6 Conclusion and Future Work

In this paper, we presented a novel unsupervised neural alignment model with contrastive learning objective, named MirrorAlign, that has achieved better alignment performance compared to FastAlign and other neural alignment models while preseving training efficiency. We empirically and theoretical show its effectiveness and reasonableness over several language pairs.

In future work, we would further explore the relationship between CLWEs and word alignments. A promising attempt is using our model as a bridge to unify cross-lingual word embeddings and word alignment tasks.

References

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. On the alignment problem in multi-head attention-based neural machine translation. In *WMT*, 2018.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

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

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. Named entity recognition with bilingual constraints. In *NAACL*, 2013.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. Incorporating structural alignment biases into an attentional neural translation model. In *NAACL*, 2016.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 2011.

Liang Ding, Longyue Wang, and Dacheng Tao. Self-attention with cross-lingual position representation. In *ACL*, 2020.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. Saliency-driven word alignment interpretation for neural machine translation. *WMT*, 2019.

Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*, 2013.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. Jointly learning to align and translate with transformer models. In *EMNLP*, 2019.

Jonas Gehring, Michael Auli, David Grangier, et al. Convolutional sequence to sequence learning. In *ICML*, 2017.

Hamidreza Ghader and Christof Monz. What does attention in neural machine translation pay attention to? In *IJCNLP*, 2017.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *ICML*, 2010.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 2012.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *EMNLP*, 2013.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *WNMT*, 2017.

Mikhail Kozhevnikov and Ivan Titov. Cross-lingual transfer of semantic role labeling models. In *ACL*, 2013.

Joël Legrand, Michael Auli, and Ronan Collobert. Neural network-based word alignment through score aggregation. In *WMT*, 2016.

Tomer Levinboim, Ashish Vaswani, and David Chiang. Model invertibility regularization: Sequence alignment with or without parallel data. In *NAACL*, 2015.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. On the word alignment from neural machine translation. In *ACL*, 2019.

Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *NAACL*, 2006.

Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015.

Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *NAACL*, 2003.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *ACL*, 2000.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. A discriminative neural model for cross-lingual word alignment. In *EMNLP*, 2019a.

Elias Stengel-Eskin, Tzu-Ray Su, Matt Post, and Benjamin Van Durme. A discriminative neural model for cross-lingual word alignment. In *EMNLP*, 2019b.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NeurIPS*, 2014.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, 2013.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. Recurrent neural networks for word alignment model. In *ACL*, 2014.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. Understanding neural machine translation by simplification: The case of encoder-free models. In *arXiv*, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

David Vilar, Maja Popović, and Hermann Ney. Aer: Do we need to ‘‘improve‘‘ our alignments? In *IWSLT*, 2006.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *COLING*, 1996.

David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT*, 2001.

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

Thomas Zenkel, Joern Wuebker, and John DeNero. End-to-end neural word alignment outperforms GIZA++. In *ACL*, 2020a.

Thomas Zenkel, Joern Wuebker, and John DeNero. End-to-end neural word alignment outperforms giza++. In *arXiv*, 2020b.