Multilingual Sentence Transformer as A Multilingual Word Aligner

Weikang Wang[1*] **Guanhua Chen**[2*] **Hanqing Wang**[1] **Yue Han**[1] **Yun Chen**[1†]
[1]Shanghai University of Finance and Economics
[2]Southern University of Science and Technology
wwk@163.sufe.edu.cn ghchen08@gmail.com
{whq,hanyue}@163.sufe.edu.cn yunchen@sufe.edu.cn

Abstract

Multilingual pretrained language models (mPLMs) have shown their effectiveness in multilingual word alignment induction. However, these methods usually start from mBERT or XLM-R. In this paper, we investigate whether multilingual sentence Transformer LaBSE is a strong multilingual word aligner. This idea is non-trivial as LaBSE is trained to learn language-agnostic sentence-level embeddings, while the alignment extraction task requires the more fine-grained word-level embeddings to be language-agnostic. We demonstrate that the vanilla LaBSE outperforms other mPLMs currently used in the alignment task, and then propose to finetune LaBSE on parallel corpus for further improvement. Experiment results on seven language pairs show that our best aligner outperforms previous state-of-the-art models of all varieties. In addition, our aligner supports different language pairs in a single model, and even achieves new state-of-the-art on zero-shot language pairs that does not appear in the finetuning process.

1 Introduction

Word alignment aims to find the correspondence between words in parallel texts (Brown et al., 1993). It is useful in a variety of natural language processing (NLP) applications such as noisy parallel corpus filtering (Kurfalı and Östling, 2019), bilingual lexicon induction (Shi et al., 2021), code-switching corpus building (Lee et al., 2019; Lin et al., 2020) and incorporating lexical constraints into neural machine translation (NMT) models (Hasler et al., 2018; Chen et al., 2021b).

Recently, neural word alignment approaches have developed rapidly and outperformed statistical word aligners like GIZA++ (Och and Ney, 2003) and fast-align (Dyer et al., 2013). Some works (Garg et al., 2019; Li et al., 2019; Zenkel et al., 2019, 2020; Chen et al., 2020b; Zhang and van Genabith, 2021; Chen et al., 2021a) induce alignments from NMT model or its variants. However, these bilingual models only support the language pair involved in the training process. They also treat the source and target side differently, thus two models are required for bidirectional alignment extraction. Another line of works (Jalili Sabet et al., 2020; Dou and Neubig, 2021) build multilingual word aligners with contextualized embeddings from the multilingual pretrained language model (Wu and Dredze, 2019; Conneau et al., 2020, mPLM). Thanks to the language-agnostic representations learned with multilingual masked language modeling task, these methods are capable of inducing word alignments even for language pairs without any parallel corpus.

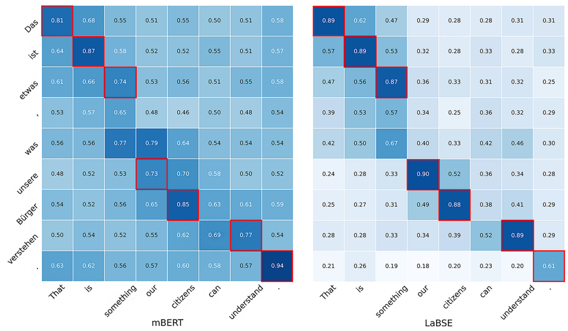

Figure 1: Cosine similarities between subword representations in a parallel sentence pair from 8th layer of mBERT (left) and 6th layer of LaBSE (right). Red boxes denote the gold alignments.

Different from previous methods, in this paper we present AccAlign, a more accurate multilingual word aligner with the multilingual sentence Transformer LaBSE (Feng et al., 2022, see Figure 1). The LaBSE is trained on large scale parallel corpus of various language pairs to learn *language-agnostic sentence embeddings* with contrastive learning. However, it is unclear whether LaBSE has learned *language-agnostic word-level embeddings*, which is the key for the success of

[*]The first two authors contribute equally.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2952–2963

word alignment extraction. Specifically, we first direct induce word alignments from LaBSE and demonstrate that LaBSE outperforms other mPLMs currently used in the alignment task. This indicates that LaBSE has *implicitly* learned language-agnostic word-level embeddings at some intermediate layer. Then we propose a simple and effective finetuning method to further improve performance. Empirical results on seven language pairs show that our best aligner outperforms previous SOTA models of all varieties. In addition, our aligner supports different language pairs in a single model, and even achieves new SOTA on zero-shot language pairs that does not appear in finetuning process.[1]

2 AccAlign

2.1 Background: LaBSE

LaBSE (Feng et al., 2022) is the state-of-the-art model for the cross-lingual sentence retrieval task. Given an input sentence, the model can retrieve the most similar sentence from candidates in a different language. LaBSE is first pretrained on a combination of masked language modeling (Devlin et al., 2019) and translation language modeling (Conneau and Lample, 2019) tasks. After that, it is effectively finetuned with contrastive loss on 6B parallel sentences across 109 languages. We leave the training detail of LaBSE in the appendix. However, as LaBSE does not include any word-level training loss when finetuning with contrastive loss, it is unclear whether the model has learned high-quality language-agnostic word-level embeddings, which is the key for a multilingual word aligner.

2.2 Alignment Induction from LaBSE

To investigate whether LaBSE is a strong multilingual word aligner, we first induce word alignments from vanilla LaBSE without any modification or finetuning. This is done by utilizing the contextual embeddings from LaBSE. Specifically, consider a bilingual sentence pair $\mathbf{x} = \langle x_1, x_2, ..., x_n \rangle$ and $\mathbf{y} = \langle y_1, x_2, ..., y_m \rangle$, we denote the contextual embeddings from LaBSE as $h_{\mathbf{x}} = \langle h_{x_1}, ..., h_{x_n} \rangle$ and $h_{\mathbf{y}} = \langle h_{y_1}, ..., h_{y_m} \rangle$, respectively. Following previous work (Dou and Neubig, 2021; Jalili Sabet et al., 2020), we get the similarity matrix from the contextual embeddings:

$$S = h_{\mathbf{x}} h_{\mathbf{y}}^T. \tag{1}$$

[1]Code is available at https://github.com/sufenlp/AccAlign.

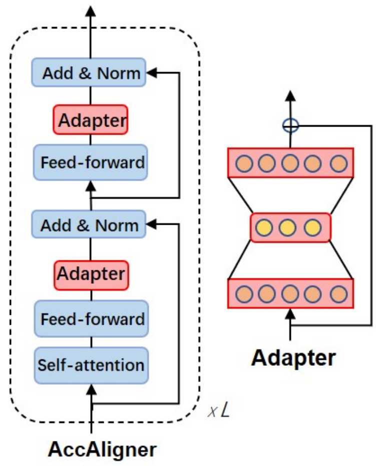

Figure 2: The framework of adapter-based finetuning. The blue blocks are kept frozen, while the red adapter blocks are updated during finetuning.

The similarity matrix is normalized for each row to get $S_{\mathbf{xy}}$. $S_{\mathbf{xy}}$ is treated as the probability matrix as its i-th row represents the probabilities of aligning x_i to all tokens in $\mathbf{y}$. The reverse probability matrix $S_{\mathbf{yx}}$ is computed similarly by normalizing each column of S. Taking intersection of the two probability matrices yields the final alignment matrix:

$$A = (S_{\mathbf{xy}} > c) * (S_{\mathbf{yx}}^T > c), \tag{2}$$

where c is a threshold and $A_{ij} = 1$ indicates that x_i and y_j are aligned. The above method induces alignments on the subword level, which are converted into word-level alignments by aligning two words if any of their subwords are aligned following (Zenkel et al., 2020; Jalili Sabet et al., 2020).

2.3 Finetuning LaBSE for Better Alignments

Inspired by (Dou and Neubig, 2021), we propose a finetuning method to further improve performance given parallel corpus with alignment labels.

Adapter-based Finetuning Adapter-based finetuning (Houlsby et al., 2019; Bapna and Firat, 2019; He et al., 2021) is not only parameter-efficient, but also benefits model performance, especially for low-resource and cross-lingual tasks (He et al., 2021). Figure 2 illustrates our overall framework, where the adapters are adopted from (Houlsby et al., 2019). For each layer of LaBSE, we introduce an adapter for each sublayer, which maps the input vector of dimension d to dimension m where $m < d$, and then re-maps it back to dimension d. Let h and h' denote the input and output vector,

Model	Setting	de-en	sv-en	fr-en	ro-en	ja-en	zh-en	fa-en	avg
				Bilingual Statistical Methods					
fast-align (Dyer et al., 2013)		27.0	-	10.5	32.1	51.1	38.1	-	-
eflomal (Östling and Tiedemann, 2016)	scratch	22.6	-	8.2	25.1	47.5	28.7	-	-
GIZA++ (Och and Ney, 2003)		20.6	-	5.9	26.4	48.0	35.1	-	-
				Bilingual Neural Methods					
MTL-FULLC-GZ (Garg et al., 2019)		16.0	-	4.6	23.1	-	-	-	-
BAO-GUIDE (Zenkel et al., 2020)		16.3	-	5.0	23.4	-	-	-	-
SHIFT-AET (Chen et al., 2020b)	scratch	15.4	-	4.7	21.2	-	17.2	-	-
MASK-ALIGN (Chen et al., 2021a)		14.4	-	4.4	19.5	-	13.8	-	-
BTBA-FCBO-SST (Zhang and van Genabith, 2021)		14.3	-	6.7	**18.5**	-	-	-	-
				Multilingual Neural Methods					
SimAlign (Jalili Sabet et al., 2020)	no ft	18.8	11.2	7.6	27.2	46.6	21.6	32.7	23.7
	no ft	17.4	9.7	5.6	27.9	45.6	18.1	33.0	22.5
AwesomeAlign (Dou and Neubig, 2021)	self-sup ft	15.9	7.9	4.4	26.2	42.4	14.9	27.1	19.8
	sup ft	15.2	7.2	4.0	25.5	40.6	13.4	25.8	18.8
	no ft	16.0	7.3	4.5	20.8	43.3	16.2	23.4	18.8
AccAlign	self-sup ft	14.3	5.8	3.9	21.6	39.2	13.0	22.6	17.2
	sup ft	**13.6**	**5.2**	**2.8**	20.8	**36.9**	**11.5**	**22.2**	**16.1**

Table 1: AER comparison between AccAlign and the baselines on test set of 7 language pairs. self-sup and sup mean finetuning the model with parallel corpus of self-supervised and human-annotated alignment labels, respectively. All multilingual methods are tested on zero-shot language pairs.

respectively. The output vector h' is calculated as:

$$h^{'} = W_{up} \cdot \tanh(W_{down} \cdot h) + h. \tag{3}$$

Note that a skip-connection is employed to approximate an identity function if parameters of the projection matrices are near zero. During finetuning, only parameters of the adapters are updated.

Training Objective Let $\hat{A}$ denote the alignment labels for the given sentence pair $\mathbf{x}$ and $\mathbf{y}$. We define the learning objective as:

$$L = \sum_{ij} \hat{A}_{ij} \frac{1}{2} \left(\frac{(S_{\mathbf{xy}})_{ij}}{n} + \frac{(S^T_{\mathbf{yx}})_{ij}}{m} \right), \tag{4}$$

where $S_{\mathbf{xy}}$ and $S_{\mathbf{yx}}$ are the alignment probability matrices, n and m are the length of sentence $\mathbf{x}$ and $\mathbf{y}$, respectively. Intuitively, this objective encourages the gold aligned words to have closer contextualized representations. In addition, as both $S_{\mathbf{xy}}$ and $S^T_{\mathbf{yx}}$ are encouraged to be close to $\hat{A}$, it implicitly encourages the two alignment probability matrices to be symmetrical to each other as well.

Our framework can be easily extended to cases where alignment labels are unavailable, by replacing $\hat{A}$ with pseudo labels A (Equation 2) and training in a self-supervised manner.

3 Experiments

3.1 Setup

As we aim at building an accurate multilingual word aligner, we evaluate AccAlign on a diverse alignment test set of seven language pairs: de/sv/ro/fr/ja/zh/fa-en. For finetuning LaBSE, we use nl/cs/hi/tr/es/pt-en as the training set and cs-en as the validation set. To reduce the alignment annotation efforts and the finetuning cost, our training set only contains $3,362$ annotated sentence pairs. To simulate the most difficult use cases where the test language pair may not included in training, we set the test language pairs different from training and validation. Namely, LaBSE is tested in a zero-shot manner. We denote this dataset as *ALIGN6*.

We induce alignments from 6-th layer of LaBSE, which is selected on the validation set. We use Alignment Error Rate (AER) as the evaluation metric. Our model is not directly comparable to the bilingual baselines, as they build model for each test language pair using large scale parallel corpus of that language pair. In contrast, our method is more efficient as it supports all language pairs in a single model and our finetuning only requires $3,362$ sentence pairs. Appendix B show more dataset, model, baselines and other setup details.

3.2 Main Results

Table 1 shows the comparison of our methods against baselines. AccAlign-supft achieves new SOTA on word alignment induction, outperforming all baselines in 6 out of 7 language pairs. AccAlign is also simpler than AwesomeAlign, which is the best existing multilingual word aligner, as AwesomeAlign finetunes with a combination of five objectives, while AccAlign only has one objective. The vanilla LaBSE is a strong multilingual word

Model		fi-el	fi-he
SimAglin	noft	69.3	85.8
AwesomeAlign	noft	69.8	84.4
	self-sup ft	68.8	87.7
	sup ft	67.4	86.1
AccAlign	noft	47.0	81.2
	self-sup ft	40.8	76.1
	sup ft	**36.7**	**71.7**

Table 2: AER comparison between AccAlign and multilingual baselines on non-English zero-shot language pairs. The best AER for each column is bold and underlined.

aligner (see AccAlign-noft). It performs better than SimAlign-noft and AwesomeAlign-noft, and comparable with AwesomeAlign-supft, indicating that LaBSE has learned high-quality language-agnostic word embeddings. Our finetuning method is effective as well, improving AccAlign-noft by 1.6 and 2.7 AER with self-supervised and supervised alignment labels, respectively. Our model improves multilingual baselines even more significantly on non-English language pairs. See Table 2 of appendix for detailed results.

3.3 Analysis

Performance on non-English Language Pair We conduct experiments to evaluate AccAlign against multilingual baselines on non-English test language pairs. The fi-el (Finnish-Greek) and fi-he (Finnish-Hebrew) test set contains 791 and 2,230 annotated sentence pairs, respectively. Both test sets are from ImaniGooghari et al. (2021)[2]. The results are shown in Table 2. As can be seen, AccAlign in all three settings significantly improves all multilingual baselines. The improvements is much larger compared with zero-shot English language pairs, demonstrating the effectiveness of AccAlign on non-English language pairs. We also observe that finetuning better improves AccAlign than AwesomeAlign. This verifies the strong cross-lingual transfer ability of LaBSE , even between English-centric and non-English language pairs.

Adapter-based vs. Full Finetuning We compare full and adapter-based fine-tuning in Table 3. Compared with full finetuning, adapter-based finetuning updates much less parameters and obtains better performance under both supervised and self-supervised settings, demonstrating its efficiency and effectiveness for zero-shot word alignments.

[2]https://github.com/cisnlp/graph-align

Ft type		full	adapter
Ft mode	self-supervised (avg.)	17.4	17.2
	supervised (avg.)	16.2	16.1
Number of ft param.		428M	2.4M

Table 3: AER comparison of full finetuning and adapter-based finetuning.

Bilingual Finetuning To better understand our method, we compare with AwesomeAlign under bilingual finetuning setup where the model is finetuned and tested in the same single language pair. We follow the setup in (Dou and Neubig, 2021) and use finetuning corpus without human-annotated labels. As shown in Table 4, LaBSE outperforms AwesomeAlign in the finetuning language pair (18.8 vs. 18.2). The performance gap becomes larger for zero-shot language pairs (21.3 vs. 18.8). The results demonstrate that AccAlign is an effective zero-shot aligner, as LaBSE has learned more language-agnostic representations which benefit cross-lingual transfer.

Different Multilingual Pretrained Models We investigate the performance of AccAlign-noft when replacing LaBSE with other mPLMs, including XLM-R, mBERT and four other multilingual sentence Transformer from HuggingFace. LaBSE outperforms other mPLMs by 3.5 to 9.6 averaged AER. Table 9 in appendix shows more details.

Performance across Layer We investigate the performance of AccAlign-noft when extracts alignments from different layers. Layer 6, which is the layer we use for all experiments, outperforms other layers by 0.1 to 26.0 averaged AER. Please refer to Table 10 in appendix for more details.

Representation Analysis To succeed in multilingual word alignment, the contextual embeddings should prefer two following properties: (1) language-agnostic: two aligned bilingual words should be mapped to nearby features in the same language-agnostic feature space. (2) word-identifiable: the embeddings of two random tokens from the same sentence should be distinguishable.

Therefore, we analyze the embeddings from different layers of AccAlign under different settings by computing cosine similarity for aligned word pairs and word pairs randomly sampled from the same sentence, denoted as s_{bi} and s_{mono} (see appendix for more experiment details). Intuitively, bigger s_{bi} and smaller s_{mono} are preferred as we

Model	Ft lang. / Test lang.	de-en	fr-en	ro-en	ja-en	zh-en	avg.
AwesomeAlign	ft lang.	14.9	4.0	22.9	38.1	14.1	18.8
	zero-shot langs (avg.)	16.3	4.7	26.6	43.7	15.0	21.3
AccAlign	ft lang.	14.2	3.8	21.0	38.0	13.8	18.2
	zero-shot langs (avg.)	14.8	3.9	20.7	40.5	13.8	18.8

Table 4: AER results with bilingual finetuning.

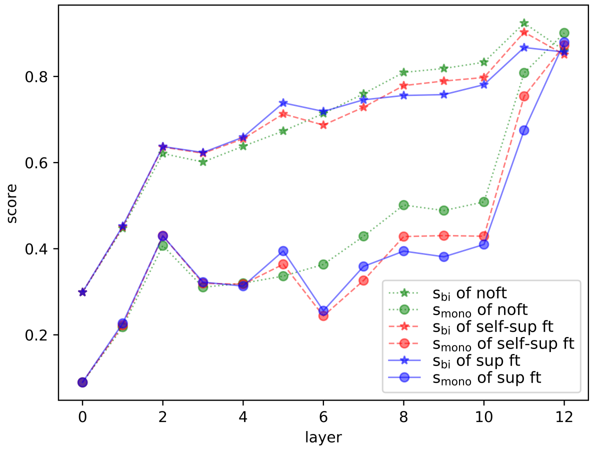

Figure 3: s_{bi} (↑) and s_{mono} (↓) of AccAlign without finetuning (noft), with self-supervised finetuning (self-sup ft) and supervised finetuning (sup ft).

expect the features of aligned words to be similar while that of two different words to be different. The results on de-en test set are presented in Figure 3. For vanilla LaBSE (green curves), we find that features from 6-th layer, namely the best layer to induce alignment, successfully trades off these two properties as it obtains the biggest $s_{\text{bi}} - s_{\text{mono}}$ among all layers. In addition, adapter-based finetuning improves performance mainly by making features more word-identifiable, as it significantly decreases s_{mono} while almost maintaining s_{bi} .

4 Conclusion

In this paper, we introduce AccAlign, a novel multilingual word aligner based on multilingual sentence Transformer LaBSE. The best proposed approach finetunes LaBSE on a few thousands of annotated parallel sentences and achieves state-of-the-art performance even for zero-shot language pairs. AccAlign is believed to be a valuable alignment tool that can be used out-of-the-box for other NLP tasks.

Limitations

AccAlign has shown to extract high quality word alignments when the input texts are two well-paired bilingual sentences. However, the condition is not always met. In lexically constrained decoding of NMT (Hasler et al., 2018; Song et al., 2020; Chen et al., 2021b), the aligner takes a full source-language sentence and a partial target-language translation as the input at each step to determine the right position to incorporate constraints. In creating translated training corpus in zero-resource language for sequence tagging or parsing (Ni et al., 2017; Jain et al., 2019; Fei et al., 2020), the aligner extracts alignments from the labelled sentence and its translation to conduct label projection. Both cases deviate from our current settings as the input sentence may contain translation error or even be incomplete. We leave exploring the robustness of AccAlign as the future work.

At the same time, our proposed method only supports languages included in LaBSE. This hinders applying AccAlign to more low-resource languages. Future explorations are needed to rapidly adapt AccAlign to new languages (Neubig and Hu, 2018; Garcia et al., 2021).

Acknowledgements

This project was supported by National Natural Science Foundation of China (No. 62106138) and Shanghai Sailing Program (No. 21YF1412100). We thank the anonymous reviewers for their insightful feedbacks on this work.

References

Niraj Aswani and Robert Gaizauskas. 2005. Aligning words in english-hindi parallel corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 115–118.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

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*, 19(2):263–311.

Mehmet Talha Cakmak, Süleyman Acar, and Gülşen Eryiğit. 2012. Word alignment for english-turkish language pair. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2177–2180.

Chi Chen, Maosong Sun, and Yang Liu. 2021a. Mask-align: Self-supervised neural word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online. Association for Computational Linguistics.

Guanhua Chen, Yun Chen, and Victor OK Li. 2021b. Lexically constrained neural machine translation with explicit alignment guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12630–12638.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020b. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. 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 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. 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 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Joao Graca, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a golden collection of parallel multi-language word alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 2 (Short Papers), pages 506–512.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Maria Holmqvist and Lars Ahrenberg. 2011. A gold standard for english-swedish word alignment. In *Proceedings of the 18th Nordic conference of computational linguistics (NODALIDA 2011)*, pages 106–113.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Ayyoob ImaniGooghari, Masoud Jalili Sabet, Lutfi Kerem Senel, Philipp Dufter, François Yvon, and Hinrich Schütze. 2021. Graph algorithms for multiparallel word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8457–8469, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092, Hong Kong, China. 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 *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2019. Noisy parallel corpus filtering through projected word embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 277–281, Florence, Italy. Association for Computational Linguistics.

Grandee Lee, Xianghu Yue, and Haizhou Li. 2019. Linguistically motivated parallel data augmentation for code-switch language modeling. In *INTERSPEECH*, pages 3730–3734.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663.

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

Lieve Macken. 2010. An annotation scheme and gold standard for dutch-english word alignment. In *7th conference on International Language Resources and Evaluation (LREC 2010)*, pages 3369–3374. European Language Resources Association (ELRA).

David Mareček. 2011. Automatic alignment of tectogrammatical trees from czech-english parallel corpus.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

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

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 813–826, Online. Association for Computational Linguistics.

Kai Song, Kun Wang, Heng Yu, Yue Zhang, Zhongqiang Huang, Weihua Luo, Xiangyu Duan, and Min Zhang. 2020. Alignment-enhanced transformer for constraining nmt with pre-specified translations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8886–8893.

Leila Tavakoli and Heshaam Faili. 2014. Phrase alignments in parallel corpus using bootstrapping approach.

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

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.

Jingyi Zhang and Josef van Genabith. 2021. A bidirectional transformer based alignment model for unsupervised word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 283–292, Online. Association for Computational Linguistics.

A LaBSE

LaBSE (Feng et al., 2022) is the state-of-the-art model for the cross-lingual sentence retrieval task. Given an input sentence, the model can retrieve the most similar sentence from candidates in a different language. It has 471M parameters and supports 109 languages. The model is first pretrained on a combination of masked language modeling (Devlin et al., 2019) and translation language modeling (Conneau and Lample, 2019) tasks on the 17B monolingual data and 6B bilingual translation pairs, respectively. After that, it is effectively finetuned with contrastive loss on 6B bilingual translation pairs across 109 languages.

Specifically, given a bilingual sentence pair $\langle \mathbf{x}^i, \mathbf{y}^i \rangle$, we use $e_{\mathbf{x}^i}$ and $e_{\mathbf{y}^i}$ to denote their sentence embeddings from LaBSE. Then the model is finetuned using contrative loss with in-batch negatives (Chen et al., 2020a):

$$\ell = -\frac{1}{N}\sum_{i=1}^{N}\left\{\log \frac{\exp\left(\phi(e_{\mathbf{x}^i}, e_{\mathbf{y}^i})\right)}{\sum_{j=1}^{N}\exp\left(\phi(e_{\mathbf{x}^i}, e_{\mathbf{y}^j})\right)} + \log \frac{\exp\left(\phi(e_{\mathbf{x}^i}, e_{\mathbf{y}^i})\right)}{\sum_{j=1}^{N}\exp\left(\phi(e_{\mathbf{x}^j}, e_{\mathbf{y}^i})\right)}\right\}, \tag{5}$$

where $\phi(e_{\mathbf{x}^i}, e_{\mathbf{y}^j})$ measures the similarity of sentence $\mathbf{x}^i$ and $\mathbf{y}^j$ in the embedding space:

$$\phi\left(e_{\mathbf{x}^i}, e_{\mathbf{y}^j}\right) = \begin{cases} e_{\mathbf{x}^i}^{\top} e_{\mathbf{y}^j} - b & \text{if } i = j \\ e_{\mathbf{x}^i}^{\top} e_{\mathbf{y}^j} & \text{if } i \neq j \end{cases}. \tag{6}$$

Note that a margin b is introduced to improve the separation between positive and negative pairs.

B Experiments Setup

B.1 Language Code

We refer to the language information in Table 1 of (Fan et al., 2021). The information of the languages used in this paper is listed in Table 5.

B.2 Dataset

Table 6 shows the detailed data statistics of ALIGN6. The ja and zh sentences are preprocessed by Dou and Neubig (2021) and Liu and Sun (2015), respectively. For finetuning AccAlign and multilingual baselines, we use the training and validation set from ALIGN6. As bilingual baselines are not capable of zero-shot alignment induction, they are trained from scratch with parallel corpus of the test language pair using the same dataset as Dou and Neubig (2021). The bilingual training data set of de/fr/ro/ja/zh-en contain 1.9M, 1.1M, 450K, 444K and 40K parallel sentence pairs, respectively, which are much larger than the training dataset of ALIGN6.

ISO	Name	Family
en	English	Germanic
nl	Dutch	Germanic
cs	Czech	Slavic
hi	Hindi	Indo-Aryan
tr	Turkish	Turkic
es	Spanish	Romance
pt	Portuguese	Romance
de	German	Germanic
sv	Swedish	Germanic
fr	French	Romance
ro	Romanian	Romance
ja	Japanese	Japonic
zh	Chinese	Chinese
fa	Persian	Iranian

Table 5: The information of the languages used in this paper.

B.3 Model Setup

We use the contextual word embeddings from the 6-th layer of the official LaBSE[3], which have 768 dimensions. We set the threshold in Equation 2 to 0.1, which is selected on validation set by manual tuning among $[0, 0.2]$. For adapter-based finetuning, we set the hidden dimension of the adapters to be 128. The adapters have 2.4M parameters, which account 0.5% of the parameters of LaBSE. We use the AdamW optimizer with learning rate of 1e-4, and do not use warmup or dropout. The batch size is set to 40 and maximum updates number is 1500 steps. We use a single NVIDIA V100 GPU for all experiments.

B.4 Baselines

Besides three statistical baselines fast-align (Dyer et al., 2013), eflomal (Östling and Tiedemann, 2016) and GIZA++ (Och and Ney, 2003), we compare AccAlign with the following neural baselines:

MTL-FULLC-GZ (Garg et al., 2019). This model supervises an attention head in Transformer-based NMT model with GIZA++ word alignments in a multitask learning framework.

BAO-GUIDE (Zenkel et al., 2020). This model

[3] https://huggingface.co/sentence-transformers/LaBSE

Type	Lang.	Source	Link	# Sents
Training set	cs-en	Mareček (2011)	http://ufal.mff.cuni.cz/czech-english-manual-word-alignment	2400
	nl-en	Macken (2010)	http://www.tst.inl.nl	372
	hi-en	Aswani and Gaizauskas (2005)	http://web.eecs.umich.edu/~mihalcea/wpt05/	90
	tr-en	Cakmak et al. (2012)	http://web.itu.edu.tr/gulsenc/resources.htm	300
	es-en	Graca et al. (2008)	https://www.hlt.inesc-id.pt/w/Word_Alignments	100
	pt-en	Graca et al. (2008)	https://www.hlt.inesc-id.pt/w/Word_Alignments	100
Validation set	cs-en	Mareček (2011)	http://ufal.mff.cuni.cz/czech-english-manual-word-alignment	101
Test set	de-en	Vilar et al. (2006)	http://www-i6.informatik.rwth-aachen.de/goldAlignment/	508
	sv-en	Holmqvist and Ahrenberg (2011)	https://www.ida.liu.se/divisions/hcs/nlplab/resources/ges/	192
	fr-en	Mihalcea and Pedersen (2003)	http://web.eecs.umich.edu/~mihalcea/wpt/	447
	ro-en	Mihalcea and Pedersen (2003)	http://web.eecs.umich.edu/~mihalcea/wpt05/	248
	ja-en	Neubig (2011)	http://www.phontron.com/kftt	582
	zh-en	Liu and Sun (2015)	https://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html	450
	fa-en	Tavakoli and Faili (2014)	http://eceold.ut.ac.ir/en/node/940	400

Table 6: Training, validation and test dataset of ALIGN6. Note that this is a zero-shot setting as the test language pairs do not appear in training and validation.

adds an extra alignment layer to repredict the to-be-aligned target token and further improves performance with Bidirectional Attention Optimization.

SHIFT-AET (Chen et al., 2020b). This model trains a separate alignment module in a self-supervised manner, and induce alignments when the to-be-aligned target token is the decoder input.

MASK-ALIGN (Chen et al., 2021a). This model is a self-supervised word aligner which makes use of the full context on the target side.

BTBA-FCBO-SST (Zhang and van Genabith, 2021). This model has similar idea with Chen et al. (2021a), but with different model architecture and training objectives.

SimAlign (Jalili Sabet et al., 2020). This model is a multilingual word aligner which induces alignment with contextual word embeddings from mBERT and XLM-R.

AwesomeAlign (Dou and Neubig, 2021). This model improves over SimAlign by designing new alignment induction method and proposing to further finetune the mPLM on parallel corpus.

Among them, SimAlign and AwesomeAlign are multilingual aligners which support multiple language pairs in a single model, while others are bilingual word aligners which require training from scratch with bilingual corpus for each test language pair. We re-implement SimAlign and AwesomeAlign, while quote the results from (Dou and Neubig, 2021) for the three statistical baselines and the corresponding paper for other baselines.

B.5 Sentence Transformer

We compare LaBSE with four other multilingual sentence Transformer in HuggingFace. The detailed information of these models are:

distiluse-base-multilingual-cased-v2.[4] This model is a multilingual knowledge distilled version of m-USE (Yang et al., 2020), which has 135M parameters and supports more than 50+ languages.

paraphrase-xlm-r-multilingual-v1.[5] This model is a multilingual version of paraphrase-distilroberta-base-v1 (Reimers and Gurevych, 2019), which has 278M parameters and supports 50+ languages. It initializes the student model with an mPLM and trains it to imitate monolingual sentence Transformer on parallel data with knowledge distillation.

paraphrase-multilingual-MiniLM-L12-v2.[6] This model is a multilingual version of paraphrase-MiniLM-L12-v2 (Reimers and Gurevych, 2019), which has 118M parameters and supports 50+ languages. It trains similarly as paraphrase-xlm-r-multilingual-v1, but with different teacher and student model initialization.

paraphrase-multilingual-mpnet-base-v2.[7] This model is a multilingual version of paraphrase-mpnet-base-v2 (Reimers and Gurevych, 2019),

[4] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

[5] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

[6] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[7] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

which has 278M parameters and supports 50+ languages. It trains similarly as paraphrase-xlm-r-multilingual-v1, but with different teacher model initialization.

B.6 Bilingual Finetuning

We use the same dataset as bilingual baselines for bilingual finetuning following (Dou and Neubig, 2021). At each time, we finetune LaBSE with one language pair among de/fr/ro/ja/zh-en and test on all seven language pairs. For Awesome-align, we follow the setup in their paper, while for AccAlign, we use the same hyperparameters as the main experiments.

B.7 Representation Analysis

We conduct representation analysis on de-en test set. To compute s_{bi}, we calculate the averaged cosine similarity of all gold aligned bilingual word pairs. To compute s_{mono}, we randomly permute a given sentence $\mathbf{x} = \langle x_1, x_2, ..., x_n \rangle$ to get $\mathbf{x}' = \langle x'_1, x'_2, ..., x'_n \rangle$ and then create n word pairs as $\{\langle x_i\text{-}x'_i \rangle\}_{i=1}^{n}$. We go through all de and en test sentences and report the averaged cosine similarity of all created word pairs as s_{mono}.

C Experiment Results

Detailed results for each test language in Section 3.3 are shown in Table 7 to Table 10.

Ft mode	Ft type	de-en	sv-en	fr-en	ro-en	ja-en	zh-en	fa-en	avg
Self-supervised	full	14.7	5.8	3.7	21.6	39.9	13.3	22.7	17.4
	adapter	14.3	5.8	3.9	21.6	39.2	13.0	22.6	17.2
Supervised	full	**13.6**	5.3	2.8	21.0	37.1	**11.0**	22.5	16.2
	adapter	**13.6**	**5.2**	**2.7**	**20.8**	**36.8**	11.5	**22.2**	**16.1**

Table 7: AER comparison of full finetuning and adapter-based finetuning. The best AER for each column is bold and underlined.

Model	Ft lang. \ Test lang.	de-en	fr-en	ro-en	ja-en	zh-en	sv-en	fa-en
AwesomeAlign	de-en	**14.9**	4.7	26.2	43.6	14.6	7.1	28.2
	fr-en	16.4	**4.0**	26.9	44.6	15.7	7.6	28.0
	ro-en	15.8	4.7	**22.9**	44.2	15.1	7.8	27.0
	ja-en	16.8	4.9	27.0	**38.1**	15.2	8.5	30.0
	zh-en	16.2	4.6	26.2	42.4	**14.1**	8.1	28.0
AccAlign	de-en	**14.2**	3.8	20.9	39.3	13.1	5.7	22.5
	fr-en	14.6	**3.8**	20.8	41.0	14.1	6.0	22.5
	ro-en	15.2	4.0	**21.0**	42.1	14.4	6.5	23.2
	ja-en	14.8	3.9	20.3	**38.0**	13.5	6.3	22.5
	zh-en	14.6	3.9	20.7	38.9	**13.4**	5.9	22.4

Table 8: AER results with bilingual finetuning. The results where the model is trained and tested on the same language pair are bold and underlined.

	layer	de-en	sv-en	fr-en	ro-en	ja-en	zh-en	fas-en	avg
mBERT	8	17.4	8.7	5.6	27.9	45.6	18.1	33.0	22.3
XLM-R	8	23.1	13.3	9.2	28.6	62.0	30.3	28.6	27.9
distiluse-base-multilingual-cased-v2	3	23.7	17.2	9.8	29.2	56.3	29.2	33.5	28.4
paraphrase-xlm-r-multilingual-v1	6	17.4	8.7	4.9	24.7	53.8	26.1	26.5	23.2
paraphrase-multilingual-MiniLM-L12-v2	6	19.4	9.4	6.2	26.0	57.7	29.7	27.4	25.1
paraphrase-multilingual-mpnet-base-v2	5	18.0	8.9	5.4	24.1	54.9	25.7	25.5	23.2
LaBSE	6	**16.0**	**7.3**	**4.5**	**20.8**	**43.3**	**16.2**	**23.4**	**18.8**

Table 9: AER comparison of LaBSE and other multilingual pretrained model. All are without finetuning. We determine the best layer of alignment induction for each model using the validation set. The best AER for each column is bold and underlined.

Layer	de-en	sv-en	fr-en	ro-en	ja-en	zh-en	fa-en	avg
0	32.4	27.7	20.5	44.2	65.5	40.1	38.7	38.4
1	27.3	19.7	12.8	35.6	64.0	33.9	35.4	32.7
2	22.3	14.0	8.6	28.8	58.0	25.0	31.3	26.9
3	18.5	9.9	6.0	24.0	50.3	17.9	26.8	21.9
4	17.7	8.7	5.9	23.3	48.4	16.3	25.7	20.9
5	**15.8**	7.4	**4.5**	21.5	43.7	15.4	23.8	18.9
6	16.0	**7.3**	**4.5**	**20.8**	43.3	16.2	23.4	**18.8**
7	16.5	7.6	4.8	22.4	43.4	**15.0**	23.7	19.1
8	16.2	**7.3**	5.0	21.6	**42.7**	16.7	23.4	19.0
9	16.8	7.6	5.3	21.5	**42.7**	17.9	**23.2**	19.3
10	17.7	9.0	5.6	23.0	44.4	20.4	24.4	20.6
11	36.7	27.0	24.2	43.6	61.3	35.0	46.2	39.1
12	43.1	33.2	30.5	46.0	65.7	42.6	52.4	44.8

Table 10: AER comparison of vanilla LaBSE across layers. Layer 0 is the embedding layer. The best AER for each column is bold and underlined.