Explicit Alignment Objectives for Multilingual Bidirectional Encoders

Junjie Hu[1]*, **Melvin Johnson**[2], **Orhan Firat**[2], **Aditya Siddhant**[2], **Graham Neubig**[1]
[1]Carnegie Mellon University, [2]Google Research
{junjieh,gneubig}@cs.cmu.edu,{melvinp,orhanf,adisid}@google.com

Abstract

Pre-trained cross-lingual encoders such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a) have proven impressively effective at enabling transfer-learning of NLP systems from high-resource languages to low-resource languages. This success comes despite the fact that there is no explicit objective to align the contextual embeddings of words/sentences with similar meanings across languages together in the same space. In this paper, we present a new method for learning multilingual encoders, AMBER (**A**ligned **M**ultilingual **B**idirectional **E**ncode**R**). AMBER is trained on additional parallel data using two *explicit* alignment objectives that align the multilingual representations at different granularities. We conduct experiments on zero-shot cross-lingual transfer learning for different tasks including sequence tagging, sentence retrieval and sentence classification. Experimental results on the tasks in the XTREME benchmark (Hu et al., 2020) show that AMBER obtains gains of up to 1.1 average F1 score on sequence tagging and up to 27.3 average accuracy on retrieval over the XLM-R-large model which has 3.2x the parameters of AMBER. Our code and models are available at http://github.com/junjiehu/amber.

1 Introduction

Cross-lingual embeddings, both traditional non-contextualized word embeddings (Faruqui and Dyer, 2014) and the more recent contextualized word embeddings (Devlin et al., 2019), are an essential tool for cross-lingual transfer in downstream applications. In particular, multilingual contextualized word representations have proven effective in reducing the amount of supervision needed in a variety of cross-lingual NLP tasks such as sequence labeling (Pires et al., 2019), question answering (Artetxe et al., 2020), parsing (Wang et al., 2019), sentence classification (Wu and Dredze, 2019) and retrieval (Yang et al., 2019a).

Some attempts at training multilingual representations (Devlin et al., 2019; Conneau et al., 2020a) simply train a (masked) language model on monolingual data from many languages. These methods can only *implicitly* learn which words and structures correspond to each-other across languages in an entirely unsupervised fashion, but are nonetheless quite effective empirically (Conneau et al., 2020b; K et al., 2020). On the other hand, some methods directly leverage multilingual parallel corpora (McCann et al., 2017; Eriguchi et al., 2018; Conneau and Lample, 2019; Huang et al., 2019; Siddhant et al., 2020), which gives some degree of supervision implicitly aligning the words in the two languages. However, the pressure on the model to learn clear correspondences between the contextualized representations in the two languages is still implicit and somewhat weak. Because of this, several follow-up works (Schuster et al., 2019; Wang et al., 2020; Cao et al., 2020) have proposed methods that use word alignments from parallel corpora as the supervision signals to align multilingual contextualized representations, albeit in a *post-hoc* fashion.

In this work, we propose a training regimen for learning contextualized word representations that encourages symmetry at both the word and sentence levels *at training time*. Our word-level alignment objective is inspired by work in machine translation that defines objectives encouraging consistency between the source-to-target and target-to-source attention matrices (Cohn et al., 2016). Our sentence-level alignment objective encourages prediction of the correct translations within a mini-batch for a given source sentence, which is inspired by work on learning multilingual sentence representations (Yang et al., 2019a; Wieting et al., 2019). In experiments, we evaluate the zero-shot cross-lingual transfer performance of AMBER on four dif-

*Work partially done at Google Research.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3633–3643
June 6–11, 2021.

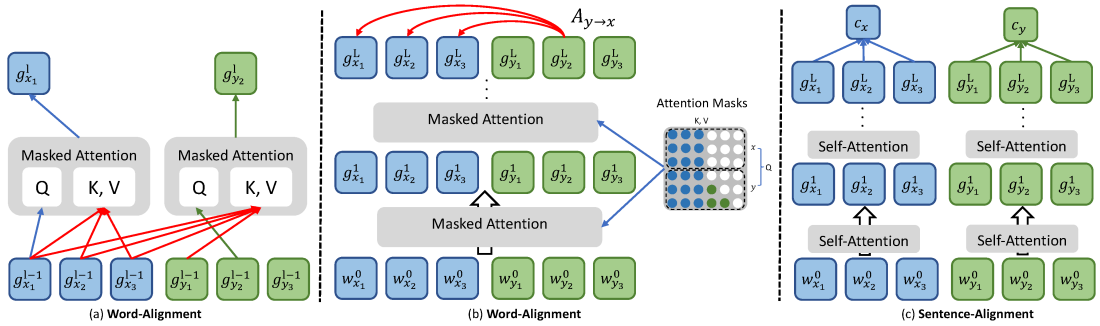


Figure 1: (a-b) show the computation of the target-to-source attention matrix used for the **word alignment** objective: (a) Masked attention for source/target (blue/green) sentences on the l-th layer; (b) Attention from y to x on the top layer. (c) shows the separate encoding of source/target sentences for the **sentence alignment** objective.

ferent NLP tasks in the XTREME benchmark (Hu et al., 2020) including part-of-speech (POS) tagging, paraphrase classification, and sentence retrieval. We show that AMBER obtains gains of up to 1.1 average F1 score on cross-lingual POS tagging, up to 27.3 average accuracy score on sentence retrieval, and achieves competitive accuracy in paraphrase classification when compared with the XLM-R-large model. This is despite the fact that XLM-R-large is trained on data 23.8x as large[1] and has 3.2x parameters of AMBER. This shows that compared to large amounts of monolingual data, even a small amount of parallel data leads to significantly better cross-lingual transfer learning.

2 Cross-lingual Alignment

This section describes three objectives for training contextualized embeddings. We denote the monolingual and parallel data as $\mathcal{M}$ and $\mathcal{P}$ respectively.

Masked Language Modeling (MLM) A masked language modeling objective takes a pair of sentences x, y, and optimizes the prediction of randomly masked tokens in the concatenation of the sentence pair as follows:

$$\ell_{\text{MLM}}(x, y) = -\mathbb{E}_{s\sim[1,|z|]} \log P(z_s|z_{\setminus s}), \quad (1)$$

where z is the concatenation of the sentence pair $z = [x; y]$, z_s are the masked tokens randomly sampled from z, and $z_{\setminus s}$ indicates all the other tokens except the masked ones.

In the standard monolingual setting, x, y are two contiguous sentences in a monolingual corpus. In Conneau and Lample (2019), x, y are two sentences in different languages from a parallel corpus, an objective we will refer to as Translation Language Modeling (TLM).

Sentence Alignment Our first proposed objective encourages cross-lingual alignment of sentence representations. For a source-target sentence pair (x, y) in the parallel corpus, we separately calculate sentence embeddings denoted as $\mathbf{c}_x, \mathbf{c}_y$ by averaging the embeddings in the final layer as the sentence embeddings.[2] We then encourage the model to predict the correct translation y given a source sentence x. To do so, we model the conditional probability of a candidate sentence y being the correct translation of a source sentence x as:

$$P(y|x) = \frac{e^{\mathbf{c}_x^T \mathbf{c}_y}}{\sum_{y'\in\mathcal{M}\cup\mathcal{P}} e^{\mathbf{c}_x^T \mathbf{c}_{y'}}}. \quad (2)$$

where y' can be any sentence in any language. Since the normalization term in Eq. (2) is intractable, we approximate $P(y|x)$ by sampling y' within a mini-batch $\mathcal{B}$ rather than $\mathcal{M} \cup \mathcal{P}$. We then define the sentence alignment loss as the average negative log-likelihood of the above probability:

$$\ell_{\text{SA}}(x, y) = -\log P(y|x). \quad (3)$$

Bidirectional Word Alignment Our second proposed objective encourages alignment of word embeddings by leveraging the attention mechanism in the Transformer model. Motivated by the work on encouraging the consistency between the source-to-target and target-to-source translations (Cohn et al., 2016; He et al., 2016), we create two different attention masks as the inputs to the Transformer model, and obtain two attention matrices in the top layer of the Transformer model. We compute the target-to-source attention matrix $\mathbf{A}_{y\to x}$ as follows:

[1] AMBER is trained on 26GB parallel data and 80GB monolingual Wikipedia data, while XLM-R-large is trained on 2.5TB monolingual CommonCrawl data.

[2] In comparison, mBERT encodes a sentence pair jointly, then uses the CLS token embedding to perform its next sentence prediction task.

$$\mathbf{g}^l_{y_i} = \text{Attn}(\mathbf{Q} = \mathbf{g}^{l-1}_{y_i}, \mathbf{KV} = \mathbf{g}^{l-1}_{[y_{<i};x]}; W^l), \quad (4)$$

$$\mathbf{g}^l_{x_j} = \text{Attn}(\mathbf{Q} = \mathbf{g}^{l-1}_{x_j}, \mathbf{KV} = \mathbf{g}^{l-1}_{x}; W^l), \quad (5)$$

$$\text{Attn}(\mathbf{QKV}; W) = \text{softmax}(\mathbf{QW}^q(\mathbf{KW}^k)^T)\mathbf{VW}^v \quad (6)$$

$$A_{y\to x}[i,j] = \mathbf{g}^L_{y_i} \cdot \mathbf{g}^L_{x_j}. \quad (7)$$

where $\mathbf{g}^l_{y_t}$ is the embedding of the t-th word in y on the l-th layer, $A_{y\to x}[i,j]$ is the (i,j)-th value in the attention matrix from y to x, and $W = \{\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v\}$ are the linear projection weights for Q, K, V respectively. We compute the source-to-target matrix $\mathbf{A}_{x\to y}$ by switching x and y.

To encourage the model to align source and target words in both directions, we aim to minimize the distance between the forward and backward attention matrices. Similarly to Cohn et al. (2016), we aim to maximize the trace of two attention matrices, i.e., $\text{tr}\left(\mathbf{A}_{y\to x}{}^T \mathbf{A}_{x\to y}\right)$. Since the attention scores are normalized in $[0,1]$, the trace of two attention matrices is upper bounded by $\min(|x|,|y|)$, and the maximum value is obtained when the two matrices are identical. Since the Transformer generates multiple attention heads, we average the trace of the bidirectional attention matrices generated by all the heads denoted by the superscript h

$$\ell_{\text{WA}}(x,y) = 1 - \frac{1}{H}\sum_{h=1}^{H} \frac{\text{tr}\left(\mathbf{A}^h_{y\to x}{}^T \mathbf{A}^h_{x\to y}\right)}{\min(|x|,|y|)}. \quad (8)$$

Notably, in the target-to-source attention in Eq (4), with attention masking we enforce a constraint that the t-th token in y can only perform attention over its preceding tokens $y_{<t}$ and the source tokens in x. This is particularly useful to control the information access of the query token y_t, in a manner similar to that of the decoding stage of NMT. Without attention masking, the standard Transformer performs self-attention over all tokens, i.e., $Q = K = \mathbf{g}^h_z$, and minimizing the distance between the two attention matrices by Eq. (8) might lead to a trivial solution where $\mathbf{W}^q \approx \mathbf{W}^k$.

Combined Objective Finally we combine the masked language modeling objective with the alignment objectives and obtain the total loss in Eq. (9). Notice that in each iteration, we sample a mini-batch of sentence pairs from $\mathcal{M} \cup \mathcal{P}$.

$$\mathcal{L} = \mathbb{E}_{(x,y)\in\mathcal{M}\cup\mathcal{P}}\, \ell_{\text{MLM}}(x,y) + \mathbb{E}_{(x,y)\in\mathcal{P}}\left[\ell_{\text{SA}}(x,y) + \ell_{\text{WA}}(x,y)\right] \quad (9)$$

Model	Data	Langs	Vocab	Layers	Parameters	Ratio
AMBER	Wiki & MT	104	120K	12	172M	1.0
mBERT	Wiki	104	120K	12	172M	1.0
XLM-15	Wiki & MT	15	95K	12	250M	1.5x
XLM-100	Wiki	100	200K	12	570M	3.3x
XLM-R-base	CommonCrawl	100	250K	12	270M	1.6x
XLM-R-large	CommonCrawl	100	250K	24	550M	3.2x
Unicoder	CommonCrawl & MT	100	250K	12	270M	1.6x

Table 1: Details of baseline and state-of-the-art models.

3 Experiments

3.1 Training setup

Following the setting of Hu et al. (2020), we focus on *zero-shot cross-lingual transfer* where we fine-tune models on English annotations and apply the models to predict on non-English data.

Models: Table 1 shows details of models in comparison. We adopt the same architecture as mBERT for AMBER. Notably, AMBER, XLM-15 and Unicoder are trained on the additional parallel data, while the others are trained only on monolingual data. Besides, XLM-R-base/large models have 2.6x/4.8x the parameters of AMBER and are trained on the larger CommonCrawl corpus. We use a simple setting for our AMBER variants in the ablation study to show the effectiveness of our proposed alignment objectives without other confounding factors such as model sizes, hyper-parameters and tokenizations in different existing studies.

Pre-training: We train AMBER on the Wikipedia data for 1M steps first using the default hyper-parameters as mBERT[3] except that we use a larger batch of 8,192 sentence pairs, as this has proven effective in Liu et al. (2019). We then continue training the model by our objectives for another 1M steps with a batch of 2,048 sentence pairs from Wikipedia corpus and parallel corpus which is used to train XLM-15 (Conneau and Lample, 2019). We use the same monolingual data as mBERT and follow Conneau and Lample (2019) to prepare the parallel data with one change to maintain truecasing. We set the maximum number of subwords in the concatenation of each sentence pair to 256 and use 10k warmup steps with the peak learning rate of 1e-4 and a linear decay of the learning rate. We train AMBER on TPU v3 for about 1 week.

3.2 Datasets

Cross-lingual Part-Of-Speech (POS) contains data in 13 languages from the Universal Dependencies v2.3 (Nivre et al., 2018).

[3] https://github.com/google-research/bert

PAWS-X (Yang et al., 2019b) is a paraphrase detection dataset. We train on the English data (Zhang et al., 2019), and evaluate the prediction accuracy on the test set translated into 4 other languages.

XNLI (Conneau et al., 2018) is a natural language inference dataset in 15 languages. We train models on the English MultiNLI training data (Williams et al., 2018), and evaluate on the other 14.

Tatoeba (Artetxe and Schwenk, 2019) is a testbed for parallel sentence identification. We select the 14 non-English languages covered by our parallel data, and follow the setup in Hu et al. (2020) finding the English translation for a given a non-English sentence with maximum cosine similarity.

3.3 Result Analysis

In Table 2, we show the average results over all languages in all the tasks, and show detailed results for each language in Appendix A.3. First, we find that our re-trained mBERT (AMBER with MLM) performs better than the publicly available mBERT on all the tasks, confirming the utility of pre-training BERT models with larger batches for more steps (Liu et al., 2019). Second, AMBER trained by the word alignment objective obtains a comparable average F1 score with respect to the best performing model (Unicoder) in the POS tagging task, which shows the effectiveness of the word-level alignment in the syntactic structure prediction tasks at the token level. Besides, it is worth noting that Unicoder is initialized from the larger XLM-R-base model that is pre-trained on a larger corpus than AMBER, and Unicoder improves over XLM-R-base on all tasks. Third, for the sentence classification tasks, AMBER trained with our explicit alignment objectives obtain a larger gain (up to 2.1 average accuracy score in PAWS-X, and 3.9 average accuracy score in XNLI) than AMBER with only the MLM objective. Although we find that AMBER trained with only the MLM objective falls behind existing XLM/XLM-R/Unicoder models with many more parameters, AMBER trained with our alignment objectives significantly narrows the gap of classification accuracy with respect to XLM/XLM-R/Unicoder. Finally, for sentence retrieval tasks, we find that XLM-15 and Unicoder are both trained on additional parallel data, outperforming the other existing models trained only on monolingual data. Using additional parallel data, AMBER with MLM and TLM objectives also significantly improves over AMBER with the MLM objective by 15.6 average accuracy score, while combining our word-level alignment objective yields a marginal improvement over AMBER with MLM and TLM objectives. However, adding the sentence-level alignment objective, AMBER trained by the combined objective can further improve AMBER with the MLM and word-level alignment objectives by 19.1 average accuracy score. This confirms our intuition that the explicit sentence-level objective can effectively leverage the alignment supervision in the parallel corpus, and encourage contextualized sentence representations of aligned pairs to be close according to the cosine similarity metric.

Model	POS	PAWS-X	XNLI	Tatoeba
mBERT (public)	68.5	86.2	65.4	45.6
XLM-15	68.8	88.0	72.6	**77.2**
XLM-100	69.5	86.4	69.1	36.6
XLM-R-base	68.8	87.4	73.4	57.6
XLM-R-large	70.0	**89.4**	**79.2**	60.6
Unicoder	**71.7**	88.1	74.8	72.2
AMBER (MLM)	69.8	87.1	67.7	52.6
AMBER (MLM+TLM)	70.5	87.7	70.9	68.2
AMBER (MLM+TLM+WA)	**71.1**	89.0	71.3	68.8
AMBER (MLM+TLM+WA+SA)	70.5	**89.2**	**71.6**	**87.9**

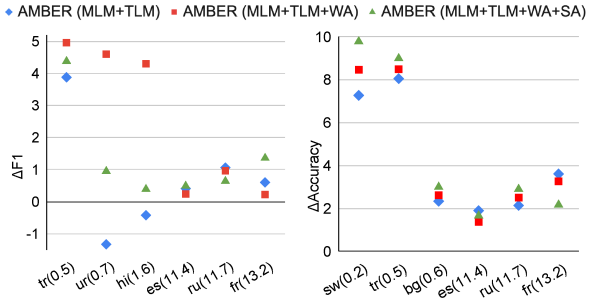
Table 2: Overall results on POS, PAWS-X, XNLI, Tatoeba tasks. Bold numbers highlight the highest scores across languages on the existing models (upper part) and AMBER variants (bottom part).

3.4 How does alignment help by language?

In Figure 2, we investigate the improvement of the alignment objectives over the MLM objective on low-resourced and high-resourced languages, by computing the performance difference between AMBER trained with alignment objectives and AMBER (MLM). First, we find that AMBER trained with alignment objectives significantly improves the performance on languages with relatively small amounts of parallel data, such as Turkish, Urdu, Swahili, while the improvement on high-resourced languages is marginal. Through a further analysis (Appendix A.3), we observe that AMBER (MLM) performs worse on these low-resourced and morphologically rich languages than on high-resourced Indo-European languages, while AMBER trained with alignment objectives can effectively bridge the gap. Moreover, AMBER trained with our word-level alignment objective yields the highest improvement on these low-resourced languages on the POS task, and AMBER trained with sentence-level alignment performs the best on XNLI.

Methods	en	bg	de	el	es	fr	Avg.
Cao et al. (2020)	80.1	73.4	73.1	71.4	75.5	74.5	74.7
AMBER (full)	84.7	74.3	74.2	72.5	76.9	76.6	76.5

Table 3: F1 scores of AMBER trained with all objectives and Cao et al. (2020) on 6 languages on XNLI.



(a) Part-of-Speech (b) XNLI

Figure 2: Performance difference between AMBER trained with alignments on parallel data and AMBER (MLM). Languages are sorted by no. of parallel data (Million) used for training AMBER with alignments.

3.5 Alignment with Attention vs Dictionary

Recent studies (Cao et al., 2020; Wang et al., 2020) have proposed to use a bilingual dictionary to align cross-lingual word representations. Compared with these methods, our word-level alignment objective encourages the model to automatically discover word alignment patterns from the parallel corpus in an end-to-end training process, which avoids potential errors accumulated in separate steps of the pipeline. Furthermore, an existing dictionary may not have all the translations for source words, especially for words with multiple senses. Even if the dictionary is relatively complete, it also requires a heuristic way to find the corresponding substrings in the parallel sentences for alignment. If we use a word alignment tool to extract a bilingual dictionary in a pipeline, errors may accumulate, hurting the accuracy of the model. Besides, Wang et al. (2020) is limited in aligning only fixed contextual embeddings from the model's top layer. Finally, we also compare AMBER trained with all the objectives and Cao et al. (2020) on a subset of languages on XNLI in Table 3. We find that our full model obtains a gain of 1.8 average F1 score.

4 Related Work

While cross-lingual alignment is a long-standing challenge dating back to the early stage of research in word alignment (Brown et al., 1993), cross-lingual embeddings (Faruqui and Dyer, 2014; Xing et al., 2015; Devlin et al., 2019; Conneau et al., 2020a) are highly promising in their easy integration into neural network models for a variety of cross-lingual applications. Analysis studies on recent cross-lingual contextualized representations (Pires et al., 2019; Wu and Dredze, 2019; Hu et al., 2020; Siddhant et al., 2020) further demonstrates this advantage for zero-shot cross-lingual transfer in a representative set of languages and tasks. In particular to improve cross-lingual transfer, some attempts directly leverage multilingual parallel corpus to train contextualized representations (McCann et al., 2017; Eriguchi et al., 2018; Conneau and Lample, 2019; Huang et al., 2019) with the hope of aligning words implicitly. The other line of work uses word alignments from parallel corpora as the alignment supervision in a post-hoc fashion (Cao et al., 2020; Wang et al., 2020). Notably, AMBER does not rely on any word alignment tools, and explicitly encourage the correspondence both on the word and sentence level.

5 Discussion and Future Work

In this paper, we demonstrate the effectiveness of our proposed explicit alignment objectives in learning better cross-lingual representations for downstream tasks. Nonetheless, several challenging and promising directions can be considered in the future. First, most existing multilingual models tokenize words into subword units, which makes the alignment less interpretable. How to align a span of subword units with meaningful semantics at the phrase level deserves further investigation. Second, several studies (Ghader and Monz, 2017; Li et al., 2019) have shown that attention may fail to capture word alignment for some language pairs, and a few works (Legrand et al., 2016; Alkhouli et al., 2018) proposed neural word alignment to improve the word alignment quality. Incorporating such recent advances into the alignment objective is one future direction. Third, how to fine-tune a well-aligned multilingual model on English annotations without catastrophic forgetting of the alignment information is a potential way to improve cross-lingual generalization on the downstream applications.

Acknowledgements

We'd like to thank Yinfei Yang and Wei-Cheng Chang for answering our questions on the data and code. JH and GN were supported in part by a Google Faculty Award, and NSF Award #1761548.

References

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. 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. 2020a. 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. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494,

Hong Kong, China. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Joël Legrand, Michael Auli, and Ronan Collobert. 2016. Neural network-based word alignment through score aggregation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 66–73, Berlin, Germany. Association for Computational Linguistics.

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.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. 2018. Universal dependencies 2.2.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8854–8861. AAAI Press.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. 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 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin soft-

max. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5370–5378. International Joint Conferences on Artificial Intelligence Organization.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019b. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendices

A.1 Training Details for Reproducibility

Although English is not the best source language for some target languages (Lin et al., 2019), this zero-shot cross-lingual transfer setting is still practical useful as many NLP tasks only have English annotations. In the following paragraphs, we show details for reproducing our results on zero-shot cross-lingual transfer setting.

Model: We use the same architecture as mBERT for AMBER, and we build our AMBER trained with the alignment objectives on top of the original mBERT implementation at https://github.com/google-research/bert, and are released at http://github.com/junjiehu/amber.

Pre-training: We first train the model on the Wikipedia data for 1M steps using the default hyper-parameters in the original repository except that we use a larger batch of 8,192 sentence pairs. The max number of subwords in the concatenation of each sentence pair is set to 256. To continue training AMBER with additional objectives on parallel data, we use 10K warmup steps with the peak learning rate of 1e-4, and use a linear decay of the learning rate. All models are pre-trained with our proposed objectives on TPU v3, and we use the same hyper-parameter setting for our AMBER variants in the experiments. We follow the practice of mBERT at https://github.com/google-research/bert/blob/master/multilingual.md#data-source-and-sampling to sample from mutlilingual data for training. We select the checkpoint of all models at the 1M step for a fair comparison. It takes about 1 week to finish the pre-training.

Fine-tuning: For fine-tuning the models on the downstream applications, we use the constant learning rate of 2e-5 as suggested in the original paper (Devlin et al., 2019). We fine-tune all the models for 10 epochs on the cross-lingual POS tag prediction task, and 5 epochs on the sentence classification task. We use the batch size of 32 for all the models. All models are fine-tuned on 2080Ti GPUs, and the training can be finished within 1 day.

Datasets: We use the same parallel data that is used to train XLM-15. The parallel data can be processed by this script: https://github.com/facebookresearch/XLM/blob/master/get-data-para.sh. All the datasets in the downstream applications can be downloaded by the script at https://github.com/google-research/xtreme/blob/master/scripts/download_data.sh. Table 4 lists all the data statistic of parallel data by languages.

A.2 Source-to-target attention matrix

We derive the source-to-target attention matrix as follow:

$$\mathbf{g}^{l}_{x_j} = \text{Attn}(\mathbf{Q} = \mathbf{g}^{l-1}_{x_j}, \mathbf{KV} = \mathbf{g}^{l-1}_{[x_{<j};y]}; W^l), \quad (10)$$

$$\mathbf{g}^{l}_{y_j} = \text{Attn}(\mathbf{Q} = \mathbf{g}^{l-1}_{y_i}, \mathbf{KV} = \mathbf{g}^{l-1}_{y}; W^l), \quad (11)$$

$$\text{Attn}(\mathbf{QKV}; W) = \text{softmax}(\mathbf{QW}^q(\mathbf{KW}^k)^T)\mathbf{VW}^v \quad (12)$$

$$A_{x\rightarrow y}[j, i] = \mathbf{g}^{L}_{x_j} \cdot \mathbf{g}^{L}_{y_i}. \quad (13)$$

A.3 Detailed Results

We show the detailed results over all languages on the cross-lingual POS task in Table 6, on the PAWS-X task in Table 5, on the XNLI task in Table 7, and on the Tatoeba retrieval task in Table 8.

A.4 Detailed Results on Performance Difference by Languages

Figure 4 and Figure 3 show the performance deference between AMBER trained with alignment objectives and AMBER trained with only MLM objective on the POS and XNLI tasks over all languages.

Language	ISO 639-1 code	# Parallel sentences (in millions)	# Wikipedia articles (in millions)	Script	Language family	Diacritics / special characters	Extensive compound-ing	Bound words / clitics	Inflec-tion	Deriva-tion
Arabic	ar	9.8	1.02	Arabic	Afro-Asiatic	X		X	X	
Bulgarian	bg	0.6	0.26	Cyrillic	IE: Slavic	X		X	X	
English	en	40.2	5.98	Latin	IE: Germanic					
French	fr	13.2	2.16	Latin	IE: Romance	X		X		
German	de	9.3	2.37	Latin	IE: Germanic		X		X	
Greek	el	4.0	0.17	Greek	IE: Greek	X	X		X	
Hindi	hi	1.6	0.13	Devanagari	IE: Indo-Aryan	X	X	X	X	X
Mandarin	zh	9.6	1.09	Chinese ideograms	Sino-Tibetan		X			
Russian	ru	11.7	1.58	Cyrillic	IE: Slavic				X	
Spanish	es	11.4	1.56	Latin	IE: Romance	X		X		
Swahili	sw	0.2	0.05	Latin	Niger-Congo			X	X	X
Thai	th	3.3	0.13	Brahmic	Kra-Dai	X				
Turkish	tr	0.5	0.34	Latin	Turkic	X	X		X	X
Urdu	ur	0.7	0.15	Perso-Arabic	IE: Indo-Aryan	X	X	X	X	X
Vietnamese	vi	3.5	1.24	Latin	Austro-Asiatic	X				

Table 4: Statistics about languages used for pre-training with our alignment objectives. Languages belong to 7 language families, with Indo-European (IE) having the most members. Diacritics / special characters: Language adds diacritics (additional symbols to letters). Compounding: Language makes extensive use of word compounds. Bound words / clitics: Function words attach to other words. Inflection: Words are inflected to represent grammatical meaning (e.g. case marking). Derivation: A single token can represent entire phrases or sentences.

Model	de	en	es	fr	zh	Avg
mBERT (public)	85.7	94.0	87.4	87.0	77.0	86.2
XLM-15	88.5	**94.7**	89.3	89.6	78.1	88.0
XLM-100	85.9	94.0	88.3	87.4	76.5	86.4
XLM-R-base	87.0	94.2	88.6	88.7	78.5	87.4
XLM-R-large	**89.7**	**94.7**	**90.1**	**90.4**	**82.3**	**89.4**
AMBER (MLM, our mBERT)	87.3	93.9	87.5	87.8	78.8	87.1
AMBER (MLM+TLM)	87.6	**95.8**	87.4	88.9	78.7	87.7
AMBER (MLM+TLM+WA)	88.9	95.5	88.9	**90.7**	**81.1**	89.0
AMBER (MLM+TLM+WA+SA)	**89.4**	95.6	**89.2**	**90.7**	80.9	**89.2**

Table 5: Accuracy of zero-shot cross-lingual classification on PAWS-X. Bold numbers highlight the highest scores across languages on the existing models (upper part) and AMBER variants (bottom part).

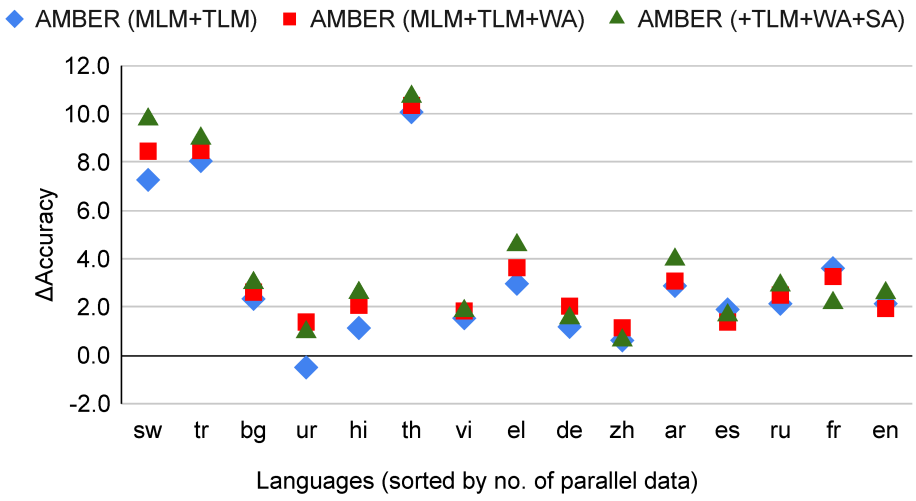


Figure 3: Performance difference between AMBER trained with alignments on parallel data and AMBER (MLM) on XNLI task. Languages are sorted by no. of parallel data used for training AMBER with alignments.

models	ar	bg	de	el	en	es	fr	hi	ru	tr	ur	vi	zh	Avg
mBERT (public)	14.9	85.2	89.3	82.8	95.3	85.7	84.1	65.1	86.0	67.5	57.4	18.5	58.9	68.5
XLM-15	17.5	86.1	89.3	85.4	95.7	85.9	84.9	63.9	86.8	69.3	55.1	18.0	57.2	68.8
XLM-100	17.1	85.8	89.3	85.7	95.4	85.3	84.3	67.0	87.1	65.0	62.0	19.2	**60.2**	69.5
XLM-R-base	17.6	**88.5**	91.1	**88.2**	95.8	87.2	85.7	70.1	88.9	72.7	61.6	19.2	27.9	68.8
XLM-R-large	**18.1**	87.4	**91.9**	87.9	**96.3**	**87.8**	**87.3**	**76.1**	**89.9**	**74.3**	**67.6**	**19.5**	26.5	**70.0**
AMBER (MLM, our mBERT)	15.4	86.6	90.1	84.3	95.5	86.5	84.6	68.2	86.8	69.0	59.2	18.7	62.1	69.8
AMBER (MLM+TLM)	**16.0**	**87.2**	**91.5**	**86.4**	**95.7**	86.9	85.2	67.7	**87.9**	72.9	57.9	19.1	**62.7**	70.5
AMBER (MLM+TLM+WA)	14.8	86.9	90.4	84.9	95.6	86.7	84.8	**72.5**	87.8	**73.9**	**63.8**	**19.5**	62.3	**71.1**
AMBER (MLM+TLM+WA+SA)	14.6	87.1	90.6	85.9	95.5	**87.0**	**86.0**	68.6	87.4	73.4	60.2	18.8	61.8	70.5

Table 6: F1 scores of part-of-speech tag predictions from the Universal Dependency v2.3. Bold numbers highlight the highest scores across languages on the existing models (upper part) and AMBER variants (bottom part).

Models	en	zh	es	de	ar	ur	ru	bg	el	fr	hi	sw	th	tr	vi	avg
mBERT (public)	80.8	67.8	73.5	70.0	64.3	57.2	67.8	68.0	65.3	73.4	58.9	49.7	54.1	60.9	69.3	65.4
XLM-15	84.1	68.8	77.8	75.7	70.4	62.2	75.0	75.7	73.3	78.0	67.3	67.5	70.5	70.0	73.0	72.6
XLM-100	82.8	70.2	75.5	72.7	66.0	59.8	69.9	71.9	70.4	74.3	62.5	58.1	65.5	66.4	70.7	69.1
XLM-R-base	83.9	73.6	78.3	75.2	71.9	65.4	75.1	76.7	75.4	77.4	69.1	62.2	72.0	70.9	74.0	73.4
XLM-R-large	**88.7**	**78.2**	**83.7**	**82.5**	**77.2**	**71.7**	**79.1**	**83.0**	**80.8**	**82.2**	**75.6**	**71.2**	**77.4**	**78.0**	**79.3**	**79.2**
AMBER (MLM, our mBERT)	82.1	71.0	75.3	72.7	66.2	60.1	70.4	71.3	67.9	74.4	63.6	50.1	55.0	64.2	71.6	67.7
AMBER (MLM+TLM)	84.3	71.6	**77.2**	73.9	69.1	59.6	72.5	73.6	70.9	**78.0**	64.7	57.4	65.0	72.2	73.1	70.9
AMBER (MLM+TLM+WA)	84.1	**72.1**	76.6	**74.7**	69.3	**61.5**	72.9	73.9	71.6	77.7	65.7	58.6	65.3	72.7	**73.4**	71.3
AMBER (MLM+TLM+WA+SA)	**84.7**	71.6	76.9	74.2	**70.2**	61.0	**73.3**	**74.3**	**72.5**	76.6	**66.2**	**59.9**	**65.7**	**73.2**	**73.4**	**71.6**

Table 7: Accuracy of zero-shot crosslingual classification on the XNLI dataset. Bold numbers highlight the highest scores across languages on the existing models (upper part) and AMBER variants (bottom part).

Method	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
mBERT (public)	25.8	49.3	77.2	29.8	68.7	66.3	34.8	61.2	11.5	13.7	34.8	31.6	62.0	71.6	45.6
XLM-15	**63.5**	71.5	**92.6**	**73.1**	**85.5**	**82.5**	**81.0**	**82.0**	**47.9**	**90.3**	**67.6**	**68.4**	**91.1**	**84.1**	**77.2**
XLM-100	18.2	40.0	66.2	25.6	58.4	54.5	26.5	44.8	12.6	31.8	26.2	18.1	47.1	42.2	36.6
XLM-R-base	36.8	67.6	89.9	53.7	74.0	74.1	54.2	72.5	19.0	38.3	61.1	36.6	68.4	60.8	57.6
XLM-R-large	47.5	**71.6**	88.8	61.8	75.7	73.7	72.2	74.1	20.3	29.4	65.7	24.3	74.7	68.3	60.6
AMBER (MLM, our mBERT)	30.7	54.9	81.4	37.7	72.7	72.7	47.5	67.5	15.1	25.7	48.3	42.6	64.6	75.1	52.6
AMBER (MLM+TLM)	47.1	61.8	89.0	53.8	76.3	77.9	72.3	69.8	20.5	83.4	88.1	50.0	86.9	78.0	68.2
AMBER (MLM+TLM+WA)	46.8	63.3	88.8	52.2	78.3	79.5	66.9	71.6	27.4	77.2	86.9	56.5	86.5	81.6	68.8
AMBER (MLM+TLM+WA+SA)	**78.5**	**87.1**	**95.5**	**75.3**	**93.3**	**92.2**	**95.0**	**91.5**	**52.8**	**94.5**	**98.4**	**84.5**	**97.4**	**94.3**	**87.9**

Table 8: Sentence retrieval accuracy on the Tatoeba dataset. Bold numbers highlight the highest scores across languages on the existing models (upper part) and AMBER variants (bottom part).

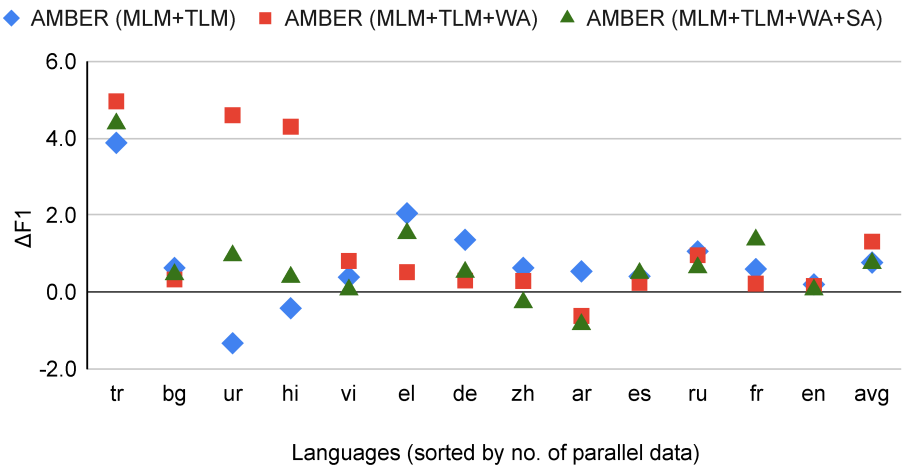


Figure 4: Performance difference between AMBER trained with alignments on parallel data and AMBER (MLM) on POS task. Languages are sorted by no. of parallel data used for training AMBER with alignments.