Do Explicit Alignments Robustly Improve Multilingual Encoders?

Shijie Wu and **Mark Dredze**
Department of Computer Science
Johns Hopkins University
shijie.wu@jhu.edu, mdredze@cs.jhu.edu

Abstract

Multilingual BERT (Devlin et al., 2019, mBERT), XLM-RoBERTa (Conneau et al., 2019, XLMR) and other unsupervised multilingual encoders can effectively learn cross-lingual representation. Explicit alignment objectives based on bitexts like Europarl or MultiUN have been shown to further improve these representations. However, word-level alignments are often suboptimal and such bitexts are unavailable for many languages. In this paper, we propose a new contrastive alignment objective that can better utilize such signal, and examine whether these previous alignment methods can be adapted to noisier sources of aligned data: a randomly sampled 1 million pair subset of the OPUS collection. Additionally, rather than report results on a single dataset with a single model run, we report the mean and standard derivation of multiple runs with different seeds, on four datasets and tasks. Our more extensive analysis finds that, while our new objective outperforms previous work, overall these methods do not improve performance with a more robust evaluation framework. Furthermore, the gains from using a better underlying model eclipse any benefits from alignment training. These negative results dictate more care in evaluating these methods and suggest limitations in applying explicit alignment objectives.

1 Introduction

Unsupervised massively multilingual encoders including multilingual BERT (Devlin et al., 2019, mBERT) and XLM-RoBERTa (Conneau et al., 2019, XLMR) are now standard tools for zero-shot cross-lingual transfer for NLP tasks (Wu and Dredze, 2019; Xia et al., 2020). While almost all encoders are pretrained without explicit cross-lingual objective, i.e. enforcing similar words from different languages have similar representation, improvements can be attained through the use of explicit cross-lingually linked data during pretraining, such as bitexts (Conneau and Lample, 2019; Huang et al., 2019; Ji et al., 2019) and dictionaries (Wu et al., 2019). As with cross-lingual embeddings (Ruder et al., 2019), these data can be used to support explicit alignment objectives with either linear mappings (Wang et al., 2019, 2020; Wu et al., 2019; Liu et al., 2019) or fine-tuning (Cao et al., 2020).

However, as word-level alignments from an unsupervised aligner are often suboptimal, we develop a new cross-lingual alignment objective for training our model. We base on our objective on contrastive learning, in which two similar inputs – such as from a bitext – are directly optimized to be similar, relative to a negative set. These methods have been effective in computer vision tasks (He et al., 2019; Chen et al., 2020a). Additionally, most previous work on contextual alignments consider high-quality bitext like Europarl (Koehn, 2005) or MultiUN (Eisele and Chen, 2010). While helpful, these resources are unavailable for most languages for which we seek a zero-shot transfer. To better reflect the quality of bitext available for most languages, we additionally use OPUS-100 (Zhang et al., 2020), a randomly sampled 1 million subset (per language pair) of the OPUS collection (Tiedemann, 2012).

We show that our new contrastive learning alignment objectives outperform previous work (Cao et al., 2020) when applied to bitext from previous works or the OPUS-100 bitext. However, our experiments also produce a negative result. While previous work showed improvements from alignment-based objectives on zero-shot cross-lingual transfer for a single task (XNLI) with a single random seed, our more extensive analysis tells a different story. We report the mean and standard derivation of multiple runs with the same hyperparam-

Code is available at https://github.com/shijie-wu/crosslingual-nlp.

eters and different random seeds. We find that previously reported improvements disappear, even while our new method shows a small improvement. Furthermore, we extend the evaluation to multiple languages on 4 tasks, further supporting our conclusions. Finally, we evaluate $\text{XLMR}_{\text{large}}$ on these tasks, which dominate the results obtained from the alignment objectives. We conclude that explicit alignments do not improve cross-lingual representations under a more extensive evaluation with noisier bitexts, and improvements are lost when compared to larger models. This negative result shows the limitation of explicit alignment objective with larger-scale bitext and encoders.

2 Explicit Alignment Objectives

We begin with a presentation of objective functions that use parallel data across languages for training multilingual encoders. These objectives assume multilingual data in the form of word pairs in parallel sentences. Since gold word alignments are scarce, we use an unsupervised word aligner. Let $\mathbf{S}$ and $\mathbf{T}$ be the contextual hidden state matrix of corresponding words from a pretrained multilingual encoder. We assume $\mathbf{S}$ is English while $\mathbf{T}$ is a combination of different target languages. As both mBERT and XLMR operate at the subword level, we use the representation of the first subword, which is consistent with the evaluation stage. Each s_i and t_i are a corresponding row of $\mathbf{S}$ and $\mathbf{T}$, respectively. $\mathbf{S}$ and $\mathbf{T}$ come from the final layer of the encoder while $\mathbf{S}^l$ and $\mathbf{T}^l$ come from the l^{th}-layer.

Linear Mapping If $\mathbf{S}$ and $\mathbf{T}$ are static feature (such as from ELMo (Peters et al., 2018)) then $\mathbf{T}$ can be aligned so that it is close to $\mathbf{S}$ via a linear mapping (Wang et al., 2019, 2020; Wu et al., 2019; Liu et al., 2019), similar to aligning monolingual embeddings to produce cross-lingual embeddings. For feature $\mathbf{S}^l$ and $\mathbf{T}^l$ from layer l, we can learn a mapping $\mathbf{W}^l$.

$$\mathbf{W}^{l*} = \arg\min_{\mathbf{W}^l} \|\mathbf{S}^l - \mathbf{T}^l\mathbf{W}^l\|_2^2 \qquad (1)$$

When $\mathbf{W}^l$ is orthogonal, Eq. (1) is known as Procrustes problem (Smith et al., 2017) and can be solved by SVD. Alternatively, Eq. (1) can also be solved by gradient descent, without the need to store in memory huge matrices $\mathbf{S}$ and $\mathbf{T}$. We adopt the latter more memory efficient approach. Following Lample et al. (2018), we enforce the orthogonality by alternating the gradient update and the following update rule

$$\mathbf{W} \leftarrow (1+\beta)\mathbf{W} - \beta(\mathbf{W}\mathbf{W}^T)\mathbf{W} \qquad (2)$$

with $\beta = 0.01$. Note we learn different $\mathbf{W}^l$ for each target language.

This approach has yielded improvements in several studies. Wang et al. (2019) used mBERT and 10k parallel sentences from Europarl to improve dependency parsing. Wang et al. (2020) used mBERT and 30k parallel sentences from Europarl to improve named entity recognition (NER) on Spanish, Dutch, and German. Wu et al. (2019) used bilingual BERT and 10k parallel sentences from XNLI (Conneau et al., 2018) to improve dependency parsing (but not NER) on French, Russian, and Chinese. Liu et al. (2019) did not evaluate on cross-lingual transfer tasks.

L2 Alignment Instead of using $\mathbf{S}$ and $\mathbf{T}$ as static features, Cao et al. (2020) proposed fine-tuning the entire encoder

$$\mathcal{L}_{\text{L2}}(\theta) = \text{mean}_i(\|s_i - t_i\|_2^2) \qquad (3)$$

where θ is the encoder parameters. To prevent a degenerative solution, they additionally use a regularization term

$$\mathcal{L}_{\text{reg-hidden}}(\theta) = \|\bar{\mathbf{S}} - \bar{\mathbf{S}}_{\text{pretrained}}\|_2^2 \qquad (4)$$

where $\bar{\mathbf{S}}$ denote **all** hidden states of the source sentence including unaligned words, encouraging the source hidden states to stay close to the pretrained hidden states. With mBERT and 20k to 250k parallel sentences from Europarl and MultiUN, Cao et al. show improvement on XNLI but not parsing.[1]

In preliminary experiments, we found constraining parameters to stay close to their original pretrained values also prevents degenerative solutions

$$\mathcal{L}_{\text{reg-param}}(\theta) = \|\theta - \theta_{\text{pretrained}}\|_2^2 \qquad (5)$$

while being more efficient than Eq. (4). As a result, we adopt the following objective (with $\lambda = 1$):

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{L2}}(\theta) + \lambda\mathcal{L}_{\text{reg-param}}(\theta) \qquad (6)$$

2.1 Contrastive Alignment

Inspired by the contrastive learning framework of Chen et al. (2020a), we propose a contrastive loss to align $\mathbf{S}$ and $\mathbf{T}$ by fine-tuning the encoder. Assume in each batch, we have corresponding (s_i, t_i)

[1] The authors state they did not observe improvements on parsing in the NLP Hightlights podcast (#112) (AI2, 2020).

where $i \in \{1, \ldots, B\}$. Instead of optimizing the absolute distance between s_i and t_i like Eq. (1) or Eq. (3), contrastive loss allows more flexibility by encouraging s_i and t_i to be closer as compared with any other hidden state. In other words, our proposed contrastive alignment optimizes the relative distance between s_i and t_i. As the alignment signal is often suboptimal, our alignment objective is more robust to errors in unsupervised word-level alignment. Additionally, unlike previous works, we select different sets of negative examples to enforce different levels of cross-lingual alignment. Finally, it naturally scales to multiple languages.

Weak alignment When the negative examples only come from target languages, we enforce a weak cross-lingual alignment, i.e. s_i should be closer to t_i than any other $t_j, \forall j \neq i$. The same is true in the other direction. The loss of a batch is

$$\begin{aligned}
&\mathcal{L}_{\text{weak}}(\theta) \\
&= \frac{1}{2B} \sum_{i=1}^{B} (\log \frac{\exp(\text{sim}(s_i, t_i)/T)}{\sum_{j=1}^{B} \exp(\text{sim}(s_i, t_j)/T)} \\
&\qquad + \log \frac{\exp(\text{sim}(s_i, t_i)/T)}{\sum_{j=1}^{B} \exp(\text{sim}(s_j, t_i)/T)}) \quad (7)
\end{aligned}$$

where $T = 0.1$ is a temperature hyperparameter and $\text{sim}(a, b)$ measures the similarity of a and b.

We use a learned cosine similarity $\text{sim}(a, b) = \cos(f(a), f(b))$ where f is a feed-forward feature extractor with one hidden layer (768-768-128) and ReLU. It can learn to discard language-specific information and only align the align-able information. Chen et al. (2020a) find that this similarity measure learns better representation for computer vision. After alignment, f is discarded as most cross-lingual transfer tasks do not need this feature extractor, though tasks like parallel sentence retrieval might find it helpful. This learned similarity cannot be applied to an absolute distance objective like Eq. (3) as it can produce degenerate solutions.

Strong alignment If the negative examples include both source and target languages, we enforce a strong cross-lingual alignment, i.e. s_i should be closer to t_i than any other $t_j, \forall j \neq i$ and $s_j, \forall j \neq i$.

$$\begin{aligned}
&\mathcal{L}_{\text{strong}}(\theta) \\
&= \frac{1}{2B} \sum_{h \in \mathcal{H}} \log \frac{\exp(\text{sim}(h, \text{aligned}(h))/T)}{\sum_{h' \in \mathcal{H}, h' \neq h} \exp(\text{sim}(h, h')/T)} \quad (8)
\end{aligned}$$

where $\text{aligned}(h)$ is the aligned hidden state of h and $\mathcal{H} = \{s_1, \ldots, s_B, t_1, \ldots, t_B\}$.

For both weak and strong alignment objectives, we add a regularization term Eq. (5) with $\lambda = 1$.

3 Experiments

Multilingual Alignment We consider alignment and transfer from English to 8 target languages: Arabic, German, English, Spanish, French, Hindi, Russian, Vietnamese, and Chinese. We use two sets of bitexts: (1) bitext used in previous works (Conneau and Lample, 2019) and (2) the OPUS-100 bitext (Zhang et al., 2020). (1) For bitext used in previous works, we use MultiUN for Arabic, Spanish, French, Russian or Chinese, EUBookshop (Skadiņš et al., 2014) for German, IIT Bombay corpus (Kunchukuttan et al., 2018) for Hindi and OpenSubtitles (Lison et al., 2018) for Vietnamese. We sample 1M bitext for each target language. (2) The OPUS-100 covering 100 languages with English as the center, and sampled from the OPUS collection randomly, which better reflects the average quality of bitext for most languages. It contains 1M bitext for each target language, except Hindi (0.5M).

We tokenize the bitext with Moses (Koehn et al., 2007) and segment Chinese with Chang et al. (2008). We use `fast_align` (Dyer et al., 2013) to produce unsupervised word alignments in both direction and symmetrize with the *grow-diag-final-and* heuristic. We only keep one-to-one alignment and discard any trivial alignment where the source and target words are identical.

We train the L2, weak, and strong alignment objectives in a multilingual fashion. Each batch contains examples from all target languages. Following Devlin et al. (2019), we optimize with Adam (Kingma and Ba, 2014), learning rate `1e-4`, 128 batch size, 100k total steps ($\approx$ 2 epochs), 4k steps linear warmup and linear decay. We use 16-bit precision and train each model on a single RTX TITAN for around 18 hours. We set the maximum sequence length to 96. For linear mapping, we use a linear decay learning rate from `1e-4` to 0 in 20k steps ($\approx$ 3 epochs), and train for 3 hours for each language pairs.

Evaluation We consider zero-shot cross-lingual transfer with XNLI (Conneau et al., 2018), NER (Pan et al., 2017), POS tagging and dependency

	XNLI	NER	POS	Parsing
mBERT	$70.1_{\pm 0.8}$	$67.7_{\pm 1.3}$	$78.3_{\pm 0.5}$	$52.6_{\pm 0.4}$
+ Linear Mapping	$70.0_{\pm 0.6}$	$63.7_{\pm 1.5}$	$79.5_{\pm 0.5}$	$53.6_{\pm 0.3}$
+ L2 Align	$69.7_{\pm 0.4}$	$67.1_{\pm 1.0}$	$78.0_{\pm 1.3}$	$52.2_{\pm 0.7}$
+ Weak Align (Our)	$70.5_{\pm 0.7}$	$68.0_{\pm 1.3}$	$78.8_{\pm 0.7}$	$53.1_{\pm 0.6}$
+ Strong Align (Our)	$70.4_{\pm 0.7}$	$67.7_{\pm 1.1}$	$79.0_{\pm 0.7}$	$53.0_{\pm 0.6}$
$\text{XLMR}_{\text{base}}$	$76.4_{\pm 0.5}$	$66.4_{\pm 0.9}$	$81.2_{\pm 0.6}$	$57.3_{\pm 0.6}$
+ Linear Mapping	$73.4_{\pm 0.6}$	$54.1_{\pm 0.9}$	$81.3_{\pm 0.5}$	$55.6_{\pm 0.5}$
+ L2 Align	$75.7_{\pm 0.5}$	$65.7_{\pm 1.2}$	$81.3_{\pm 0.9}$	$56.2_{\pm 0.7}$
+ Weak Align (Our)	$76.1_{\pm 0.7}$	$66.0_{\pm 1.0}$	$81.5_{\pm 0.5}$	$57.4_{\pm 0.4}$
+ Strong Align (Our)	$76.0_{\pm 0.6}$	$66.1_{\pm 0.9}$	$81.4_{\pm 0.6}$	$57.4_{\pm 0.5}$
$\text{XLMR}_{\text{large}}$	$80.4_{\pm 0.6}$	$71.0_{\pm 1.4}$	$82.6_{\pm 0.5}$	$59.4_{\pm 0.8}$

(a) Alignment with bitext used in previous works

	XNLI	NER	POS	Parsing
mBERT	$70.1_{\pm 0.8}$	$67.7_{\pm 1.3}$	$78.3_{\pm 0.5}$	$52.6_{\pm 0.4}$
+ Linear Mapping	$70.2_{\pm 0.6}$	$63.8_{\pm 1.3}$	$80.1_{\pm 0.4}$	$53.6_{\pm 0.3}$
+ L2 Align	$70.3_{\pm 0.5}$	$67.8_{\pm 1.4}$	$78.2_{\pm 1.2}$	$52.8_{\pm 0.7}$
+ Weak Align (Our)	$70.8_{\pm 0.7}$	$67.3_{\pm 0.9}$	$78.8_{\pm 0.6}$	$52.9_{\pm 0.6}$
+ Strong Align (Our)	$70.4_{\pm 0.7}$	$67.2_{\pm 1.1}$	$79.0_{\pm 0.7}$	$53.3_{\pm 0.6}$
$\text{XLMR}_{\text{base}}$	$76.4_{\pm 0.5}$	$66.4_{\pm 0.9}$	$81.2_{\pm 0.6}$	$57.3_{\pm 0.6}$
+ Linear Mapping	$73.5_{\pm 0.5}$	$54.2_{\pm 0.8}$	$81.7_{\pm 0.6}$	$56.1_{\pm 0.4}$
+ L2 Align	$75.8_{\pm 0.5}$	$65.5_{\pm 1.2}$	$81.4_{\pm 0.8}$	$55.9_{\pm 0.6}$
+ Weak Align (Our)	$76.0_{\pm 0.4}$	$66.2_{\pm 1.2}$	$81.5_{\pm 0.5}$	$57.4_{\pm 0.5}$
+ Strong Align (Our)	$76.1_{\pm 0.4}$	$66.2_{\pm 1.0}$	$81.5_{\pm 0.6}$	$57.4_{\pm 0.5}$
$\text{XLMR}_{\text{large}}$	$80.4_{\pm 0.6}$	$71.0_{\pm 1.4}$	$82.6_{\pm 0.5}$	$59.4_{\pm 0.8}$

(b) Alignment with the OPUS-100 bitext

Table 1: Zero-shot cross-lingual transfer result, average over 9 languages. Breakdown can be found in App. B. Blue or orange indicates the mean performance is one standard derivation above or below the mean of baseline. While mBERT benefits from alignment in some cases, extra alignment does not improve XLMR.

parsing (Zeman et al., 2020).[2] We evaluate XNLI and POS tagging with accuracy (ACC), NER with span-level F1, and parsing with labeled attachment score (LAS). For the task-specific layer, we use a linear classifier for XNLI, NER, and POS tagging, and use Dozat and Manning (2017) for dependency parsing. We fine-tune all parameters on English training data and directly transfer to target languages. We optimize with Adam, learning rate `2e-5` with 10% steps linear warmup and linear decay, 5 epochs, and 32 batch size. For the linear mapping alignment, we use an ELMo-style feature-based model[3] with 4 extra Transformer layers (Vaswani et al., 2017), a CRF instead of a linear classifier for NER, and train for 20 epochs, a batch size of 128 and learning rate `1e-3` (except NER and XNLI with `1e-4`). All token level tasks use the first subword as the word representation for task-specific layers following previous work (Devlin et al., 2019; Wu and Dredze, 2019). Model selection is done on the English dev set. We report the mean and standard derivation of test performance of 5 evaluation runs with different random seeds[4] and the same hyperparameters. Additional experiments detail can be found in App. A.

4 Result

Robustness of Previous Methods With a more robust evaluation scheme and 1 million parallel sentences (4× to 100× of previously considered data), the previously proposed Linear Mapping or L2 Alignment does not consistently outperform a no alignment setting more than one standard derivation in all cases (Tab. 1). With mBERT, L2 Alignment performs comparably to no alignment on all 4 tasks (XNLI, NER, POS tagging, and parsing). Compared to no alignment, Linear Mapping performs much worse on NER, performs better on POS tagging and parsing, and performs comparably on XNLI. While previous work observes small improvements on selected languages and tasks, it likely depends on the randomness during evaluation. Based on a more comprehensive evaluation including 4 tasks and multiple seeds, the previously proposed methods do not consistently perform better than no alignment with millions of parallel sentences.

Contrastive Alignment In Tab. 1, with mBERT, both proposed contrastive alignment methods consistently perform as well as no alignment while outperforming more than 1 standard derivation on POS tagging and/or parsing. This suggests the proposed methods are more robust to suboptimal alignments. We hypothesize that learned cosine similarity and contrastive alignment allow the model to recover from suboptimal alignments. Both weak and strong alignment perform comparably. While preliminary experiments found that increasing the batch size by 1.5× does not lead to better performance, future work could consider using a memory bank to greatly increase the number of negative examples (Chen et al., 2020b), which has been shown to be beneficial for computer vision tasks.

[2] We use the following treebanks: Arabic-PADT, German-GSD, English-EWT, Spanish-GSD, French-GSD, Hindi-HDTB, Russian-GSD, Vietnamese-VTB, and Chinese-GSD.

[3] We take the weighted average of representations in all layers of the encoder.

[4] We pick 5 random seeds before the experiment and use the same seeds for each task and model.

Alignment with XLMR XLMR, trained on 2.5TB of text, has the same number of transformer layers as mBERT but larger vocabulary. It performs much better than mBERT. Therefore, we wonder if an explicit alignment objective can similarly lead to better cross-lingual representations. Unfortunately, in Tab. 1, we find all alignment methods we consider do not improve over no alignment. Compared to no alignment, Linear Mapping and L2 Alignment have worse performance in 3 out of 4 tasks (except POS tagging). In contrast to previous work, both contrastive alignment objectives perform comparably to no alignment in all 4 tasks.

Impact of Bitext Quality Even though the OPUS-100 bitext has lower quality compared to bitext used in previous works (due to its greater inclusion of bitext from various sources), it has minimum impact on each alignment method we consider. This is good news for the lower resource languages, as not all languages are covered by MultiUN or Europarl.

Model Capacity vs Alignment $\text{XLMR}_{\text{large}}$ has nearly twice the number of parameters as $\text{XLMR}_{\text{base}}$. Even trained on the same data, it performs much better than $\text{XLMR}_{\text{base}}$, with or without alignment. This suggests increasing model capacity likely leads to better cross-lingual representations than using an explicit alignment objective. Future work could tackle the curse of multilinguality (Conneau et al., 2019) by increasing the model capacity in a computationally efficient way (Pfeiffer et al., 2020).

5 Discussion

Our proposed contrastive alignment objective outperforms L2 Alignment (Cao et al., 2020) and consistently performs as well as or better than no alignment using various quality bitext on 4 NLP tasks under a comprehensive evaluation with multiple seeds. However, to our surprise, previously proposed methods do not show consistent improvement over no alignment in this setting. Therefore, we make the following recommendations for future work on cross-lingual alignment or multilingual representations: 1) Evaluations should consider average quality data, not exclusively high-quality bitext. 2) Evaluation must consider multiple NLP tasks or datasets. 3) Evaluation should report **mean and variance over multiple seeds**, not a single run. More broadly, the community must establish a robust evaluation scheme for zero-shot cross-lingual transfer as a single run with one random seed does not reflect the variance of the method (especially in a zero-shot or few-shot setting).[5] While Keung et al. (2020) advocate using oracle for model selection, we instead argue reporting the variance of test performance, following the few-shot learning literature. Additionally, no alignment methods improve XLMR and larger $\text{XLMR}_{\text{large}}$ performs much better, and raw text is easier to obtain than bitext. Therefore, scaling models to more raw text and larger capacity models may be more beneficial for producing better cross-lingual models.

Acknowledgments

This research is supported in part by ODNI, IARPA, via the BETTER Program contract #2019-19051600005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

This research is supported by the following open-source softwares: NumPy (Harris et al., 2020), PyTorch (Paszke et al., 2017), PyTorch lightning (Falcon, 2019), scikit-learn (Pedregosa et al., 2011), Transformer (Wolf et al., 2019).

References

AI2. 2020. 112 - Alignment of multilingual contextual representations, with Steven Cao. *NLP Highlights Podcast*.

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

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

[5] This includes recently compiled zero-shot cross-lingual transfer benchmarks like XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020).

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

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.

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.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

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.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

WA Falcon. 2019. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3.

Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585:357–362.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

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 generalization. *arXiv preprint arXiv:2003.11080*.

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.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Cross-lingual pre-training based transfer for zero-shot neural machine translation. *arXiv preprint arXiv:1912.01214*.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. On the evaluation of contextual embeddings for zero-shot cross-lingual transfer learning. *arXiv preprint arXiv:2004.15001*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 33–43, Hong Kong, China. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *International Conference on Learning Representations*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

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.

Patrick Xia, Shijie Wu, and B. V. Durme. 2020. Which *bert? a survey organizing contextualized encoders.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Ethan Chi, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Samson Tella, Isabelle

Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

A Additional Experiments Detail

Evaluation Detail We set the maximum sequence length to 128 during fine-tuning. For NER and POS tagging, we additionally use a sliding window of context to include subwords beyond the first 128. At test time, we use the same maximum sequence length except for parsing. At test time for parsing, we only use the first 128 words of a sentence instead of subwords to make sure we compare different models consistently. We ignore words with POS tags of `SYM` and `PUNCT` during parsing evaluation. We rewrite the `BIO` label, similar to an unbiased structure predictor, to make sure a valid span is produced during NER evaluation. As the supervision on Chinese NER is on character-level, we segment the character into word using the Stanford Word Segmenter and realign the label.

All datasets we used are publicly available: NER[6], XNLI[7][8], POS tagging and dependency parsing[9]. Data statistic can be found in Tab. 2.

	XNLI	NER	POS tagging Parsing
en-train	392703	20000	12543
en-dev	2490	10000	2002
en-test	5010	10000	2077
ar-test	5010	10000	680
de-test	5010	10000	977
es-test	5010	10000	426
fr-test	5010	10000	416
hi-test	5010	1000	1684
ru-test	5010	10000	601
vi-test	5010	10000	800
zh-test	5010	10000	500

Table 2: Number of examples.

B Breakdown of Zero-shot Cross-lingual Transfer Result

Breakdown of alignment with bitext from previous works can be found in Tab. 3 and breakdown of alignment with the OPUS-100 bitext can be found in Tab. 4.

[6] https://www.amazon.com/clouddrive/share/d3KGCRCIYwhKJF0H3eWA26hjg2ZCRhjpEQtDL70FSBN

[7] https://cims.nyu.edu/~sbowman/multinli/multinli_1.0.zip

[8] https://dl.fbaipublicfiles.com/XNLI/XNLI-1.0.zip

[9] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3226

	ar	de	en	es	fr	hi	ru	vi	zh	AVER
XNLI (Accuracy)										
mBERT	64.2$_{\pm 0.9}$	70.5$_{\pm 0.2}$	82.5$_{\pm 0.3}$	74.2$_{\pm 1.2}$	73.8$_{\pm 0.8}$	59.4$_{\pm 0.7}$	68.3$_{\pm 0.9}$	69.6$_{\pm 0.7}$	68.6$_{\pm 0.9}$	70.1$_{\pm 0.8}$
+ Linear Mapping	63.8$_{\pm 0.6}$	70.4$_{\pm 0.4}$	81.0$_{\pm 0.5}$	73.9$_{\pm 0.9}$	72.5$_{\pm 0.8}$	61.2$_{\pm 0.7}$	67.1$_{\pm 0.4}$	70.2$_{\pm 0.5}$	70.1$_{\pm 0.8}$	70.0$_{\pm 0.6}$
+ L2 Align	64.1$_{\pm 0.4}$	70.0$_{\pm 0.7}$	82.2$_{\pm 0.4}$	73.9$_{\pm 0.5}$	73.8$_{\pm 0.2}$	58.5$_{\pm 0.3}$	67.9$_{\pm 0.4}$	69.4$_{\pm 0.6}$	67.9$_{\pm 0.4}$	69.7$_{\pm 0.4}$
+ Weak Align (Our)	64.9$_{\pm 0.8}$	71.0$_{\pm 0.8}$	82.3$_{\pm 0.4}$	74.6$_{\pm 0.7}$	73.8$_{\pm 0.4}$	59.8$_{\pm 0.3}$	68.5$_{\pm 1.0}$	70.3$_{\pm 0.8}$	69.4$_{\pm 1.0}$	70.5$_{\pm 0.7}$
+ Strong Align (Our)	64.8$_{\pm 0.8}$	70.5$_{\pm 0.9}$	82.3$_{\pm 0.5}$	74.4$_{\pm 0.6}$	74.1$_{\pm 0.7}$	59.8$_{\pm 0.9}$	68.2$_{\pm 0.6}$	70.1$_{\pm 0.8}$	69.0$_{\pm 1.0}$	70.4$_{\pm 0.7}$
XLMR$_{\text{base}}$	71.8$_{\pm 0.2}$	77.3$_{\pm 0.5}$	85.1$_{\pm 0.3}$	79.3$_{\pm 0.5}$	78.8$_{\pm 0.4}$	70.3$_{\pm 0.6}$	75.9$_{\pm 0.5}$	74.8$_{\pm 0.4}$	74.1$_{\pm 0.5}$	76.4$_{\pm 0.5}$
+ Linear Mapping	69.7$_{\pm 0.6}$	74.3$_{\pm 0.3}$	82.5$_{\pm 0.6}$	76.4$_{\pm 0.5}$	75.5$_{\pm 0.4}$	67.2$_{\pm 0.9}$	73.2$_{\pm 0.3}$	72.5$_{\pm 0.5}$	68.9$_{\pm 1.2}$	73.4$_{\pm 0.6}$
+ L2 Align	71.6$_{\pm 0.8}$	76.0$_{\pm 0.5}$	84.5$_{\pm 0.5}$	78.6$_{\pm 0.3}$	77.9$_{\pm 0.3}$	69.8$_{\pm 0.7}$	75.3$_{\pm 0.3}$	74.0$_{\pm 0.4}$	73.7$_{\pm 0.7}$	75.7$_{\pm 0.5}$
+ Weak Align (Our)	71.7$_{\pm 0.7}$	76.5$_{\pm 0.6}$	84.7$_{\pm 0.6}$	78.7$_{\pm 0.6}$	78.1$_{\pm 0.7}$	70.4$_{\pm 0.9}$	75.8$_{\pm 0.6}$	74.5$_{\pm 0.5}$	74.2$_{\pm 0.7}$	76.1$_{\pm 0.7}$
+ Strong Align (Our)	71.6$_{\pm 0.5}$	76.6$_{\pm 0.4}$	84.7$_{\pm 0.5}$	79.0$_{\pm 0.4}$	78.3$_{\pm 0.3}$	70.0$_{\pm 1.0}$	75.7$_{\pm 0.7}$	74.7$_{\pm 0.4}$	73.7$_{\pm 0.8}$	76.0$_{\pm 0.6}$
XLMR$_{\text{large}}$	77.5$_{\pm 0.6}$	81.7$_{\pm 0.4}$	88.0$_{\pm 0.3}$	83.3$_{\pm 0.6}$	82.0$_{\pm 0.5}$	75.1$_{\pm 0.8}$	79.2$_{\pm 0.7}$	78.4$_{\pm 0.6}$	78.3$_{\pm 0.6}$	80.4$_{\pm 0.6}$
NER (Entity-level F1)										
mBERT	42.0$_{\pm 2.9}$	79.0$_{\pm 0.3}$	84.1$_{\pm 0.2}$	73.3$_{\pm 2.5}$	78.9$_{\pm 0.3}$	65.7$_{\pm 1.4}$	65.2$_{\pm 1.4}$	69.7$_{\pm 1.8}$	51.7$_{\pm 0.8}$	67.7$_{\pm 1.3}$
+ Linear Mapping	36.9$_{\pm 1.1}$	76.1$_{\pm 0.4}$	82.8$_{\pm 0.1}$	70.4$_{\pm 2.1}$	77.4$_{\pm 0.7}$	64.5$_{\pm 1.4}$	59.5$_{\pm 2.5}$	65.2$_{\pm 2.7}$	40.5$_{\pm 2.0}$	63.7$_{\pm 1.5}$
+ L2 Align	39.7$_{\pm 1.6}$	77.7$_{\pm 0.8}$	84.0$_{\pm 0.1}$	72.5$_{\pm 1.5}$	79.1$_{\pm 0.3}$	63.3$_{\pm 1.8}$	64.3$_{\pm 1.0}$	71.2$_{\pm 0.9}$	52.1$_{\pm 1.1}$	67.1$_{\pm 1.0}$
+ Weak Align (Our)	42.3$_{\pm 2.7}$	78.7$_{\pm 0.3}$	84.2$_{\pm 0.2}$	71.6$_{\pm 2.2}$	79.4$_{\pm 0.6}$	67.6$_{\pm 1.3}$	64.8$_{\pm 0.8}$	70.0$_{\pm 2.3}$	52.9$_{\pm 0.9}$	68.0$_{\pm 1.3}$
+ Strong Align (Our)	40.6$_{\pm 1.0}$	78.7$_{\pm 0.3}$	84.2$_{\pm 0.2}$	72.2$_{\pm 2.5}$	79.0$_{\pm 0.5}$	67.2$_{\pm 0.7}$	64.5$_{\pm 1.7}$	70.1$_{\pm 2.5}$	52.5$_{\pm 0.8}$	67.7$_{\pm 1.1}$
XLMR$_{\text{base}}$	44.0$_{\pm 1.3}$	75.0$_{\pm 0.3}$	82.2$_{\pm 0.2}$	76.0$_{\pm 2.4}$	77.6$_{\pm 0.7}$	65.7$_{\pm 0.6}$	64.1$_{\pm 0.7}$	68.0$_{\pm 1.2}$	45.1$_{\pm 0.8}$	66.4$_{\pm 0.9}$
+ Linear Mapping	30.8$_{\pm 2.1}$	69.0$_{\pm 0.6}$	78.3$_{\pm 0.3}$	59.8$_{\pm 0.5}$	67.8$_{\pm 0.7}$	57.9$_{\pm 1.5}$	48.0$_{\pm 1.0}$	54.4$_{\pm 0.5}$	21.0$_{\pm 0.9}$	54.1$_{\pm 0.9}$
+ L2 Align	44.9$_{\pm 2.1}$	74.9$_{\pm 0.6}$	82.1$_{\pm 0.3}$	75.0$_{\pm 3.1}$	77.1$_{\pm 0.6}$	65.5$_{\pm 1.3}$	63.2$_{\pm 0.3}$	66.3$_{\pm 2.2}$	42.4$_{\pm 0.7}$	65.7$_{\pm 1.2}$
+ Weak Align (Our)	45.6$_{\pm 1.4}$	75.0$_{\pm 0.5}$	82.2$_{\pm 0.2}$	74.2$_{\pm 2.4}$	77.2$_{\pm 0.8}$	65.8$_{\pm 1.1}$	63.6$_{\pm 1.1}$	67.6$_{\pm 0.7}$	42.8$_{\pm 0.6}$	66.0$_{\pm 1.0}$
+ Strong Align (Our)	45.7$_{\pm 1.7}$	75.1$_{\pm 0.6}$	82.1$_{\pm 0.3}$	73.5$_{\pm 1.7}$	77.2$_{\pm 0.6}$	65.8$_{\pm 1.7}$	63.7$_{\pm 0.5}$	68.1$_{\pm 0.8}$	43.2$_{\pm 0.4}$	66.1$_{\pm 0.9}$
XLMR$_{\text{large}}$	46.8$_{\pm 4.3}$	79.1$_{\pm 0.5}$	84.2$_{\pm 0.2}$	75.7$_{\pm 2.9}$	80.7$_{\pm 0.5}$	71.6$_{\pm 1.1}$	71.7$_{\pm 0.5}$	77.4$_{\pm 1.3}$	51.5$_{\pm 1.4}$	71.0$_{\pm 1.4}$
POS (Accuracy)										
mBERT	60.3$_{\pm 0.9}$	90.4$_{\pm 0.3}$	96.9$_{\pm 0.1}$	87.7$_{\pm 0.2}$	88.9$_{\pm 0.3}$	68.0$_{\pm 0.8}$	82.5$_{\pm 0.7}$	62.7$_{\pm 0.2}$	67.1$_{\pm 1.1}$	78.3$_{\pm 0.5}$
+ Linear Mapping	73.6$_{\pm 0.7}$	88.2$_{\pm 0.5}$	96.3$_{\pm 0.0}$	87.4$_{\pm 0.1}$	88.9$_{\pm 0.3}$	77.3$_{\pm 0.6}$	78.0$_{\pm 1.0}$	60.4$_{\pm 0.5}$	65.7$_{\pm 1.3}$	79.5$_{\pm 0.5}$
+ L2 Align	63.4$_{\pm 2.6}$	89.3$_{\pm 0.7}$	96.7$_{\pm 0.2}$	86.7$_{\pm 0.3}$	87.9$_{\pm 0.5}$	65.2$_{\pm 3.9}$	83.6$_{\pm 0.9}$	62.3$_{\pm 0.8}$	66.5$_{\pm 1.5}$	78.0$_{\pm 1.3}$
+ Weak Align (Our)	61.6$_{\pm 2.0}$	90.3$_{\pm 0.7}$	96.9$_{\pm 0.1}$	87.5$_{\pm 0.6}$	88.6$_{\pm 0.3}$	70.3$_{\pm 0.9}$	83.1$_{\pm 0.6}$	63.2$_{\pm 0.3}$	68.1$_{\pm 0.9}$	78.8$_{\pm 0.7}$
+ Strong Align (Our)	61.9$_{\pm 2.0}$	90.4$_{\pm 0.7}$	96.9$_{\pm 0.0}$	87.5$_{\pm 0.5}$	88.5$_{\pm 0.4}$	71.1$_{\pm 1.2}$	83.0$_{\pm 0.5}$	63.2$_{\pm 0.2}$	68.0$_{\pm 0.6}$	79.0$_{\pm 0.7}$
XLMR$_{\text{base}}$	70.2$_{\pm 1.6}$	91.6$_{\pm 0.3}$	97.5$_{\pm 0.0}$	88.5$_{\pm 0.2}$	89.4$_{\pm 0.3}$	71.7$_{\pm 1.3}$	86.1$_{\pm 0.3}$	64.5$_{\pm 0.5}$	71.4$_{\pm 0.5}$	81.2$_{\pm 0.6}$
+ Linear Mapping	74.3$_{\pm 1.1}$	90.7$_{\pm 0.5}$	96.9$_{\pm 0.0}$	88.2$_{\pm 0.1}$	89.3$_{\pm 0.3}$	82.1$_{\pm 0.9}$	82.7$_{\pm 0.4}$	62.6$_{\pm 0.4}$	65.3$_{\pm 1.0}$	81.3$_{\pm 0.5}$
+ L2 Align	71.1$_{\pm 1.8}$	91.4$_{\pm 0.3}$	97.4$_{\pm 0.0}$	88.2$_{\pm 0.2}$	89.0$_{\pm 0.3}$	73.0$_{\pm 3.8}$	86.6$_{\pm 0.2}$	64.4$_{\pm 0.4}$	70.8$_{\pm 0.8}$	81.3$_{\pm 0.9}$
+ Weak Align (Our)	72.8$_{\pm 0.7}$	91.1$_{\pm 0.2}$	97.4$_{\pm 0.0}$	88.3$_{\pm 0.2}$	89.2$_{\pm 0.2}$	72.4$_{\pm 1.6}$	86.4$_{\pm 0.1}$	64.7$_{\pm 0.4}$	71.6$_{\pm 1.2}$	81.5$_{\pm 0.5}$
+ Strong Align (Our)	72.5$_{\pm 0.9}$	91.1$_{\pm 0.3}$	97.4$_{\pm 0.0}$	88.3$_{\pm 0.2}$	89.1$_{\pm 0.1}$	72.0$_{\pm 2.1}$	86.4$_{\pm 0.1}$	64.8$_{\pm 0.4}$	71.4$_{\pm 1.1}$	81.4$_{\pm 0.6}$
XLMR$_{\text{large}}$	73.9$_{\pm 1.0}$	91.9$_{\pm 0.3}$	98.0$_{\pm 0.0}$	89.2$_{\pm 0.2}$	89.8$_{\pm 0.1}$	78.4$_{\pm 2.1}$	86.5$_{\pm 0.2}$	64.8$_{\pm 0.3}$	71.0$_{\pm 0.3}$	82.6$_{\pm 0.5}$
Parsing (Labeled Attachment Score)										
mBERT	28.8$_{\pm 0.4}$	67.8$_{\pm 0.5}$	79.7$_{\pm 0.1}$	69.1$_{\pm 0.1}$	73.3$_{\pm 0.2}$	31.0$_{\pm 0.5}$	60.2$_{\pm 0.6}$	33.5$_{\pm 0.5}$	29.5$_{\pm 0.4}$	52.6$_{\pm 0.4}$
+ Linear Mapping	44.1$_{\pm 0.3}$	64.4$_{\pm 0.4}$	80.5$_{\pm 0.2}$	70.2$_{\pm 0.3}$	73.9$_{\pm 0.1}$	32.2$_{\pm 0.3}$	56.7$_{\pm 0.5}$	32.1$_{\pm 0.2}$	28.1$_{\pm 0.3}$	53.6$_{\pm 0.3}$
+ L2 Align	29.6$_{\pm 1.6}$	66.9$_{\pm 0.2}$	79.2$_{\pm 0.2}$	68.2$_{\pm 0.4}$	72.5$_{\pm 0.5}$	30.8$_{\pm 1.9}$	60.0$_{\pm 0.6}$	33.3$_{\pm 0.4}$	29.5$_{\pm 0.4}$	52.2$_{\pm 0.7}$
+ Weak Align (Our)	30.7$_{\pm 0.9}$	67.6$_{\pm 0.6}$	79.8$_{\pm 0.1}$	69.7$_{\pm 0.4}$	73.6$_{\pm 0.4}$	31.2$_{\pm 0.8}$	61.3$_{\pm 0.7}$	33.5$_{\pm 0.6}$	30.5$_{\pm 0.6}$	53.1$_{\pm 0.6}$
+ Strong Align (Our)	31.2$_{\pm 1.1}$	67.5$_{\pm 0.4}$	79.8$_{\pm 0.1}$	69.4$_{\pm 0.3}$	73.4$_{\pm 0.5}$	30.7$_{\pm 1.5}$	61.3$_{\pm 0.8}$	33.5$_{\pm 0.6}$	30.0$_{\pm 0.5}$	53.0$_{\pm 0.6}$
XLMR$_{\text{base}}$	43.7$_{\pm 1.7}$	69.0$_{\pm 0.4}$	80.5$_{\pm 0.2}$	71.0$_{\pm 0.4}$	73.6$_{\pm 0.5}$	41.2$_{\pm 0.9}$	66.3$_{\pm 0.9}$	36.6$_{\pm 0.2}$	34.2$_{\pm 0.7}$	57.3$_{\pm 0.6}$
+ Linear Mapping	47.2$_{\pm 0.6}$	66.7$_{\pm 0.3}$	81.4$_{\pm 0.1}$	72.6$_{\pm 0.2}$	74.4$_{\pm 0.4}$	41.4$_{\pm 0.7}$	60.8$_{\pm 0.6}$	34.3$_{\pm 0.3}$	21.5$_{\pm 1.1}$	55.6$_{\pm 0.5}$
+ L2 Align	41.3$_{\pm 1.8}$	68.1$_{\pm 0.3}$	79.7$_{\pm 0.2}$	70.0$_{\pm 0.5}$	73.0$_{\pm 0.5}$	40.2$_{\pm 1.6}$	63.7$_{\pm 0.9}$	36.5$_{\pm 0.5}$	32.9$_{\pm 0.3}$	56.2$_{\pm 0.7}$
+ Weak Align (Our)	44.6$_{\pm 1.0}$	68.8$_{\pm 0.4}$	80.4$_{\pm 0.1}$	71.4$_{\pm 0.2}$	73.9$_{\pm 0.2}$	41.0$_{\pm 0.6}$	65.7$_{\pm 0.4}$	36.7$_{\pm 0.4}$	33.8$_{\pm 0.3}$	57.4$_{\pm 0.4}$
+ Strong Align (Our)	44.8$_{\pm 0.9}$	68.9$_{\pm 0.5}$	80.4$_{\pm 0.1}$	71.3$_{\pm 0.2}$	73.9$_{\pm 0.1}$	40.7$_{\pm 0.8}$	66.2$_{\pm 0.4}$	36.7$_{\pm 0.3}$	34.0$_{\pm 0.8}$	57.4$_{\pm 0.5}$
XLMR$_{\text{large}}$	48.2$_{\pm 1.5}$	67.8$_{\pm 0.6}$	82.6$_{\pm 0.3}$	73.9$_{\pm 0.4}$	76.4$_{\pm 0.4}$	41.8$_{\pm 2.5}$	69.6$_{\pm 0.4}$	38.9$_{\pm 0.6}$	35.4$_{\pm 0.5}$	59.4$_{\pm 0.8}$

Table 3: Zero-shot cross-lingual transfer result with bitext from previous works. Blue or orange indicates the mean performance is one standard derivation above or below the mean of baseline.

	ar	de	en	es	fr	hi	ru	vi	zh	AVER
XNLI (Accuracy)										
mBERT	64.2$_{\pm 0.9}$	70.5$_{\pm 0.2}$	82.5$_{\pm 0.3}$	74.2$_{\pm 1.2}$	73.8$_{\pm 0.8}$	59.4$_{\pm 0.7}$	68.3$_{\pm 0.9}$	69.6$_{\pm 0.7}$	68.6$_{\pm 0.9}$	70.1$_{\pm 0.8}$
+ Linear Mapping	64.1$_{\pm 0.7}$	70.0$_{\pm 0.6}$	81.0$_{\pm 0.5}$	74.1$_{\pm 0.6}$	72.9$_{\pm 0.9}$	61.8$_{\pm 0.7}$	67.4$_{\pm 0.6}$	70.2$_{\pm 0.5}$	70.2$_{\pm 0.8}$	70.2$_{\pm 0.6}$
+ L2 Align	64.3$_{\pm 0.5}$	70.7$_{\pm 1.0}$	82.5$_{\pm 0.5}$	74.3$_{\pm 0.3}$	74.0$_{\pm 0.4}$	59.3$_{\pm 0.4}$	68.6$_{\pm 0.7}$	69.7$_{\pm 0.4}$	69.1$_{\pm 0.5}$	70.3$_{\pm 0.5}$
+ Weak Align (Our)	65.1$_{\pm 0.9}$	70.9$_{\pm 0.6}$	82.6$_{\pm 0.5}$	74.9$_{\pm 0.6}$	74.1$_{\pm 0.4}$	60.3$_{\pm 0.6}$	68.9$_{\pm 0.8}$	70.6$_{\pm 0.6}$	69.6$_{\pm 1.0}$	70.8$_{\pm 0.7}$
+ Strong Align (Our)	64.7$_{\pm 0.9}$	70.8$_{\pm 0.7}$	82.4$_{\pm 0.1}$	74.5$_{\pm 0.7}$	73.9$_{\pm 0.7}$	59.6$_{\pm 0.6}$	68.5$_{\pm 1.1}$	70.4$_{\pm 0.6}$	69.1$_{\pm 1.0}$	70.4$_{\pm 0.7}$
XLMR$_{\text{base}}$	71.8$_{\pm 0.2}$	77.3$_{\pm 0.5}$	85.1$_{\pm 0.3}$	79.3$_{\pm 0.5}$	78.8$_{\pm 0.4}$	70.3$_{\pm 0.6}$	75.9$_{\pm 0.5}$	74.8$_{\pm 0.4}$	74.1$_{\pm 0.5}$	76.4$_{\pm 0.5}$
+ Linear Mapping	69.9$_{\pm 0.4}$	74.3$_{\pm 0.3}$	82.5$_{\pm 0.6}$	76.4$_{\pm 0.5}$	75.5$_{\pm 0.6}$	67.2$_{\pm 1.0}$	72.7$_{\pm 0.2}$	72.7$_{\pm 0.5}$	70.1$_{\pm 0.8}$	73.5$_{\pm 0.5}$
+ L2 Align	71.9$_{\pm 0.6}$	76.4$_{\pm 0.4}$	84.6$_{\pm 0.3}$	78.4$_{\pm 0.5}$	77.8$_{\pm 0.3}$	69.9$_{\pm 0.8}$	75.2$_{\pm 0.5}$	74.2$_{\pm 0.5}$	73.7$_{\pm 0.5}$	75.8$_{\pm 0.5}$
+ Weak Align (Our)	71.8$_{\pm 0.6}$	76.5$_{\pm 0.5}$	84.6$_{\pm 0.2}$	79.0$_{\pm 0.4}$	78.4$_{\pm 0.5}$	70.0$_{\pm 0.5}$	75.7$_{\pm 0.3}$	74.7$_{\pm 0.3}$	73.4$_{\pm 0.6}$	76.0$_{\pm 0.4}$
+ Strong Align (Our)	72.0$_{\pm 0.5}$	76.6$_{\pm 0.4}$	84.8$_{\pm 0.1}$	79.0$_{\pm 0.4}$	78.6$_{\pm 0.5}$	70.1$_{\pm 0.3}$	75.7$_{\pm 0.4}$	74.8$_{\pm 0.6}$	73.8$_{\pm 0.6}$	76.1$_{\pm 0.4}$
XLMR$_{\text{large}}$	77.5$_{\pm 0.6}$	81.7$_{\pm 0.4}$	88.0$_{\pm 0.3}$	83.3$_{\pm 0.6}$	82.0$_{\pm 0.5}$	75.1$_{\pm 0.8}$	79.2$_{\pm 0.7}$	78.4$_{\pm 0.6}$	78.3$_{\pm 0.6}$	80.4$_{\pm 0.6}$
NER (Entity-level F1)										
mBERT	42.0$_{\pm 2.9}$	79.0$_{\pm 0.3}$	84.1$_{\pm 0.2}$	73.3$_{\pm 2.5}$	78.9$_{\pm 0.3}$	65.7$_{\pm 1.4}$	65.2$_{\pm 1.4}$	69.7$_{\pm 1.8}$	51.7$_{\pm 0.8}$	67.7$_{\pm 1.3}$
+ Linear Mapping	36.9$_{\pm 0.9}$	76.2$_{\pm 0.3}$	82.8$_{\pm 0.1}$	71.2$_{\pm 1.5}$	77.4$_{\pm 0.7}$	62.4$_{\pm 2.2}$	59.6$_{\pm 2.4}$	65.4$_{\pm 2.6}$	42.3$_{\pm 1.4}$	63.8$_{\pm 1.3}$
+ L2 Align	41.3$_{\pm 3.2}$	78.2$_{\pm 1.0}$	84.1$_{\pm 0.1}$	73.4$_{\pm 2.4}$	79.7$_{\pm 0.8}$	64.9$_{\pm 1.5}$	64.9$_{\pm 1.6}$	71.8$_{\pm 0.9}$	52.4$_{\pm 1.3}$	67.8$_{\pm 1.4}$
+ Weak Align (Our)	40.3$_{\pm 1.1}$	78.7$_{\pm 0.3}$	84.0$_{\pm 0.1}$	70.7$_{\pm 2.1}$	79.0$_{\pm 0.4}$	67.2$_{\pm 1.2}$	64.9$_{\pm 1.2}$	69.1$_{\pm 0.8}$	52.0$_{\pm 1.1}$	67.3$_{\pm 0.9}$
+ Strong Align (Our)	40.7$_{\pm 1.9}$	78.3$_{\pm 0.3}$	84.2$_{\pm 0.1}$	70.0$_{\pm 2.6}$	78.8$_{\pm 0.3}$	66.7$_{\pm 1.4}$	64.8$_{\pm 0.9}$	69.5$_{\pm 1.4}$	52.1$_{\pm 0.6}$	67.2$_{\pm 1.1}$
XLMR$_{\text{base}}$	44.0$_{\pm 1.3}$	75.0$_{\pm 0.3}$	82.2$_{\pm 0.2}$	76.0$_{\pm 2.4}$	77.6$_{\pm 0.7}$	65.7$_{\pm 0.6}$	64.1$_{\pm 0.7}$	68.0$_{\pm 1.2}$	45.1$_{\pm 0.8}$	66.4$_{\pm 0.9}$
+ Linear Mapping	30.8$_{\pm 1.6}$	69.3$_{\pm 0.6}$	78.3$_{\pm 0.3}$	60.2$_{\pm 0.8}$	67.9$_{\pm 0.5}$	58.2$_{\pm 0.7}$	47.7$_{\pm 0.8}$	54.1$_{\pm 0.3}$	21.6$_{\pm 1.2}$	54.2$_{\pm 0.8}$
+ L2 Align	44.1$_{\pm 1.2}$	74.2$_{\pm 0.7}$	81.9$_{\pm 0.3}$	74.9$_{\pm 3.3}$	76.9$_{\pm 0.6}$	64.7$_{\pm 0.5}$	61.9$_{\pm 1.4}$	68.4$_{\pm 2.2}$	42.1$_{\pm 1.1}$	65.5$_{\pm 1.2}$
+ Weak Align (Our)	45.5$_{\pm 2.8}$	75.0$_{\pm 0.8}$	82.2$_{\pm 0.2}$	73.7$_{\pm 1.8}$	77.3$_{\pm 0.6}$	66.6$_{\pm 1.3}$	64.0$_{\pm 1.2}$	67.5$_{\pm 1.4}$	43.9$_{\pm 1.2}$	66.2$_{\pm 1.2}$
+ Strong Align (Our)	45.3$_{\pm 1.5}$	75.1$_{\pm 0.4}$	82.2$_{\pm 0.2}$	74.6$_{\pm 2.5}$	77.4$_{\pm 0.6}$	66.0$_{\pm 1.2}$	63.7$_{\pm 0.9}$	68.0$_{\pm 1.1}$	43.3$_{\pm 0.4}$	66.2$_{\pm 1.0}$
XLMR$_{\text{large}}$	46.8$_{\pm 4.3}$	79.1$_{\pm 0.5}$	84.2$_{\pm 0.2}$	75.7$_{\pm 2.9}$	80.7$_{\pm 0.5}$	71.6$_{\pm 1.1}$	71.7$_{\pm 0.5}$	77.4$_{\pm 1.3}$	51.5$_{\pm 1.4}$	71.0$_{\pm 1.4}$
POS (Accuracy)										
mBERT	60.3$_{\pm 0.9}$	90.4$_{\pm 0.3}$	96.9$_{\pm 0.1}$	87.7$_{\pm 0.2}$	88.9$_{\pm 0.3}$	68.0$_{\pm 0.8}$	82.5$_{\pm 0.7}$	62.7$_{\pm 0.2}$	67.1$_{\pm 1.1}$	78.3$_{\pm 0.5}$
+ Linear Mapping	76.2$_{\pm 0.5}$	91.2$_{\pm 0.1}$	96.3$_{\pm 0.0}$	87.6$_{\pm 0.1}$	89.0$_{\pm 0.2}$	74.9$_{\pm 1.1}$	80.6$_{\pm 0.3}$	60.4$_{\pm 0.5}$	64.8$_{\pm 1.3}$	80.1$_{\pm 0.4}$
+ L2 Align	62.7$_{\pm 2.9}$	89.5$_{\pm 0.8}$	96.8$_{\pm 0.1}$	87.1$_{\pm 0.3}$	88.3$_{\pm 0.2}$	65.2$_{\pm 3.7}$	83.8$_{\pm 1.0}$	62.8$_{\pm 0.5}$	67.3$_{\pm 1.1}$	78.2$_{\pm 1.2}$
+ Weak Align (Our)	61.1$_{\pm 1.3}$	90.4$_{\pm 0.8}$	96.9$_{\pm 0.0}$	87.7$_{\pm 0.5}$	88.7$_{\pm 0.3}$	70.3$_{\pm 1.2}$	83.2$_{\pm 0.6}$	63.3$_{\pm 0.3}$	68.0$_{\pm 0.5}$	78.8$_{\pm 0.6}$
+ Strong Align (Our)	61.7$_{\pm 1.7}$	90.5$_{\pm 0.7}$	96.9$_{\pm 0.0}$	87.7$_{\pm 0.6}$	88.7$_{\pm 0.4}$	70.5$_{\pm 1.0}$	83.3$_{\pm 0.7}$	63.1$_{\pm 0.3}$	68.2$_{\pm 0.8}$	79.0$_{\pm 0.7}$
XLMR$_{\text{base}}$	70.2$_{\pm 1.6}$	91.6$_{\pm 0.3}$	97.5$_{\pm 0.0}$	88.5$_{\pm 0.2}$	89.4$_{\pm 0.3}$	71.7$_{\pm 1.3}$	86.1$_{\pm 0.3}$	64.5$_{\pm 0.5}$	71.4$_{\pm 0.5}$	81.2$_{\pm 0.6}$
+ Linear Mapping	76.0$_{\pm 0.9}$	92.0$_{\pm 0.1}$	96.9$_{\pm 0.0}$	88.7$_{\pm 0.2}$	89.5$_{\pm 0.3}$	78.9$_{\pm 2.1}$	83.9$_{\pm 0.3}$	62.5$_{\pm 0.4}$	66.5$_{\pm 1.0}$	81.7$_{\pm 0.6}$
+ L2 Align	71.0$_{\pm 0.9}$	91.2$_{\pm 0.5}$	97.3$_{\pm 0.0}$	87.9$_{\pm 0.3}$	88.8$_{\pm 0.4}$	74.8$_{\pm 2.9}$	86.9$_{\pm 0.8}$	64.0$_{\pm 0.6}$	70.6$_{\pm 0.5}$	81.4$_{\pm 0.8}$
+ Weak Align (Our)	72.5$_{\pm 0.8}$	91.2$_{\pm 0.3}$	97.4$_{\pm 0.0}$	88.2$_{\pm 0.2}$	89.2$_{\pm 0.2}$	72.7$_{\pm 1.3}$	86.2$_{\pm 0.2}$	64.7$_{\pm 0.4}$	71.8$_{\pm 1.4}$	81.5$_{\pm 0.5}$
+ Strong Align (Our)	72.5$_{\pm 0.6}$	91.2$_{\pm 0.2}$	97.4$_{\pm 0.1}$	88.3$_{\pm 0.2}$	89.2$_{\pm 0.2}$	72.0$_{\pm 1.9}$	86.5$_{\pm 0.2}$	64.8$_{\pm 0.4}$	71.7$_{\pm 1.7}$	81.5$_{\pm 0.6}$
XLMR$_{\text{large}}$	73.9$_{\pm 1.0}$	91.9$_{\pm 0.3}$	98.0$_{\pm 0.0}$	89.2$_{\pm 0.2}$	89.8$_{\pm 0.1}$	78.4$_{\pm 2.1}$	86.5$_{\pm 0.2}$	64.8$_{\pm 0.3}$	71.0$_{\pm 0.3}$	82.6$_{\pm 0.5}$
Parsing (Labeled Attachment Score)										
mBERT	28.8$_{\pm 0.4}$	67.8$_{\pm 0.5}$	79.7$_{\pm 0.1}$	69.1$_{\pm 0.1}$	73.3$_{\pm 0.2}$	31.0$_{\pm 0.5}$	60.2$_{\pm 0.6}$	33.5$_{\pm 0.5}$	29.5$_{\pm 0.4}$	52.6$_{\pm 0.4}$
+ Linear Mapping	45.0$_{\pm 0.3}$	67.7$_{\pm 0.2}$	80.5$_{\pm 0.2}$	70.0$_{\pm 0.3}$	73.9$_{\pm 0.2}$	28.4$_{\pm 0.2}$	57.2$_{\pm 0.4}$	32.0$_{\pm 0.3}$	28.1$_{\pm 0.2}$	53.6$_{\pm 0.3}$
+ L2 Align	29.7$_{\pm 0.6}$	67.7$_{\pm 0.7}$	79.3$_{\pm 0.4}$	68.9$_{\pm 0.6}$	73.4$_{\pm 0.5}$	31.7$_{\pm 1.8}$	61.3$_{\pm 1.2}$	33.6$_{\pm 0.5}$	29.7$_{\pm 0.2}$	52.8$_{\pm 0.7}$
+ Weak Align (Our)	29.9$_{\pm 1.0}$	67.6$_{\pm 0.4}$	79.8$_{\pm 0.0}$	69.6$_{\pm 0.3}$	73.5$_{\pm 0.5}$	31.0$_{\pm 1.6}$	61.2$_{\pm 0.9}$	33.4$_{\pm 0.7}$	30.0$_{\pm 0.5}$	52.9$_{\pm 0.6}$
+ Strong Align (Our)	30.8$_{\pm 0.9}$	68.0$_{\pm 0.4}$	79.8$_{\pm 0.1}$	69.9$_{\pm 0.3}$	73.7$_{\pm 0.5}$	31.5$_{\pm 1.5}$	61.8$_{\pm 0.6}$	33.5$_{\pm 0.6}$	30.4$_{\pm 0.4}$	53.3$_{\pm 0.6}$
XLMR$_{\text{base}}$	43.7$_{\pm 1.7}$	69.0$_{\pm 0.4}$	80.5$_{\pm 0.2}$	71.0$_{\pm 0.4}$	73.6$_{\pm 0.5}$	41.2$_{\pm 0.9}$	66.3$_{\pm 0.9}$	36.6$_{\pm 0.2}$	34.2$_{\pm 0.7}$	57.3$_{\pm 0.6}$
+ Linear Mapping	48.0$_{\pm 0.5}$	69.2$_{\pm 0.2}$	81.4$_{\pm 0.1}$	72.4$_{\pm 0.1}$	74.8$_{\pm 0.3}$	38.8$_{\pm 0.9}$	61.8$_{\pm 0.5}$	34.2$_{\pm 0.3}$	24.2$_{\pm 0.9}$	56.1$_{\pm 0.4}$
+ L2 Align	39.4$_{\pm 0.5}$	68.0$_{\pm 0.5}$	79.9$_{\pm 0.2}$	69.9$_{\pm 0.5}$	72.8$_{\pm 0.5}$	40.2$_{\pm 1.1}$	63.8$_{\pm 0.8}$	36.4$_{\pm 0.6}$	32.3$_{\pm 0.9}$	55.9$_{\pm 0.6}$
+ Weak Align (Our)	44.5$_{\pm 1.3}$	68.7$_{\pm 0.7}$	80.4$_{\pm 0.1}$	71.3$_{\pm 0.3}$	73.8$_{\pm 0.3}$	41.4$_{\pm 0.8}$	65.7$_{\pm 0.4}$	36.7$_{\pm 0.4}$	34.0$_{\pm 0.7}$	57.4$_{\pm 0.5}$
+ Strong Align (Our)	44.9$_{\pm 1.0}$	68.8$_{\pm 0.6}$	80.4$_{\pm 0.1}$	71.2$_{\pm 0.2}$	73.8$_{\pm 0.2}$	41.1$_{\pm 0.8}$	65.9$_{\pm 0.5}$	36.6$_{\pm 0.3}$	33.9$_{\pm 0.7}$	57.4$_{\pm 0.5}$
XLMR$_{\text{large}}$	48.2$_{\pm 1.5}$	67.8$_{\pm 0.6}$	82.6$_{\pm 0.3}$	73.9$_{\pm 0.4}$	76.4$_{\pm 0.4}$	41.8$_{\pm 2.5}$	69.6$_{\pm 0.4}$	38.9$_{\pm 0.6}$	35.4$_{\pm 0.5}$	59.4$_{\pm 0.8}$

Table 4: Zero-shot cross-lingual transfer result with the OPUS-100 bitext. Blue or orange indicates the mean performance is one standard derivation above or below the mean of baseline.