TransAlign: Machine Translation Encoders are Strong Word Aligners, Too

Anonymous ACL submission

Abstract

In the absence of sizable training data for most world languages and NLP tasks, translation-based strategies such as *translate-test*—evaluating on noisy source language data translated from the target language—and *translate-train*—training on noisy target language data translated from the source language—have been established as competitive approaches for cross-lingual transfer (XLT). For token classification tasks, these strategies require *label projection*: mapping the labels from each token in the original sentence to its counterpart(s) in the translation. To this end, it is common to leverage multilingual word aligners (WAs) derived from encoder language models such as mBERT or LaBSE. Despite obvious associations between machine translation (MT) and WA, research on extracting alignments with MT models is largely limited to exploiting cross-attention in encoder-decoder architectures, yielding poor WA results. In this work, in contrast, we propose TransAlign, a novel word aligner that utilizes the encoder of a massively multilingual MT model. We show that TransAlign not only achieves strong WA performance but substantially outperforms popular WAs and state-of-the-art non-WA-based label projection methods in MT-based XLT for token classification.

1 Motivation and Background

In recent years, multilingual language models (mLMs) have been positioned as the primary tool for cross-lingual transfer (XLT). By fine-tuning on task data in a high-resource source language, mLMs can make predictions in target languages with no (zero-shot XLT) or limited (few-shot XLT) labeled examples (Wu and Dredze, 2019; Wang et al., 2019; Lauscher et al., 2020; Schmidt et al., 2022). However, for *token classification tasks* (e.g., named entity recognition), translation-based XLT strategies—where a machine translation (MT) model is used to either (1) translate the original target language instance into the (noisy) source language before inference, known as *translate-test* (`T-Test`), or (2) generate noisy target language data by translating the original source language data before training, known as *translate-train* (`T-Train`) (Hu et al., 2020; Ruder et al., 2021; Ebrahimi et al., 2022; Aggarwal et al., 2022; Artetxe et al., 2023; Ebing and Glavaš, 2024)—substantially outperform zero-shot XLT (Chen et al., 2023; García-Ferrero et al., 2023; Le et al., 2024; Parekh et al., 2024), especially for low(er)-resource languages (Ebing and Glavaš, 2025).

Translation-based XLT strategies for token classification tasks require the additional step of *label projection*: mapping the labeled spans from the original to the translated sentence. A broad body of work addressed label projection starting from task-specific (Duong et al., 2013; Ni et al., 2017; Stengel-Eskin et al., 2019; Eskander et al., 2020; Fei et al., 2020, inter alia) and evolving to task-agnostic methods (Chen et al., 2023; García-Ferrero et al., 2023; Le et al., 2024; Parekh et al., 2024). While WA-based label projection (Och and Ney, 2003; Dyer et al., 2013; Jalili Sabet et al., 2020; Dou and Neubig, 2021; Wang et al., 2022)—which project labels by aligning tokens in the original sentence to corresponding tokens in the translated sentence—served as baseline throughout, recent work has rendered it less effective than other label projection strategies such as marker-based approaches (Chen et al., 2023; García-Ferrero et al., 2023; Le et al., 2024; Parekh et al., 2024). Ebing and Glavaš (2025), however, show that WA-based label projection can perform at least on a par with these state-of-the-art projection methods as long as: (i) it is carefully designed and (ii) relies on a strong underlying WA model.

Current multilingual WAs either leverage contextualized embeddings from vanilla encoders (e.g., mBERT or XLM-R) (Jalili Sabet et al., 2020;

Dou and Neubig, 2021) or sentence encoders (e.g., LaBSE) (Wang et al., 2022). Despite, WA and MT being, intuitively, two highly related and interleaved tasks (Och and Ney, 2003; Callison-Burch et al., 2004; Koehn et al., 2007; Dyer et al., 2013), research on extracting word alignments from MT models has largely been limited to extracting alignments from the attention mechanism, yielding poor WA performance for the cross-attention of transformer-based encoder-decoder MT models (Bahdanau et al., 2015; Ghader and Monz, 2017; Ferrando and Costa-jussà, 2021).

Contributions. In this work, **(1)** we propose TransAlign, a WA that leverages (only) the encoder of NLLB (Team et al., 2022), a massively multilingual encoder-decoder MT model. Next, to its vanilla (non-fine-tuned) variant, we explore the impact of further fine-tuning TransAlign on parallel WA data. **(2)** We extensively evaluate TransAlign extrinsically on translation-based XLT for token classification on two established benchmarks covering 28 diverse languages. We find TransAlign to substantially outperform popular word aligners as well as a state-of-the-art non-WA-based label projection method. Furthermore, we evaluate TransAlign intrinsically on the word alignment task showing its strong performance, particularly on words carrying semantic meaning. **(3)** Finally, we ablate important design decisions including the encoder layer to extract the alignments from and the similarity threshold based on which an alignment is established. We will publicly release our code.

2 An MT Encoder as a Word Aligner

The task of word alignment aims at finding semantically corresponding pairs of words between a source language sentence $x = (x_1, x_2, ..., x_n)$ and target language sentence $y = (y_1, y_2, ..., y_m)$:

$$A = \{(x_i, y_j) : x_i \in x, y_j \in y\}.$$

Extracting Alignments. For TransAlign, we extract word alignments from the contextualized embeddings produced by the *encoder* of a multilingual encoder-decoder MT model. We separately feed the source language sentence x and target language sentence y through the encoder obtaining their contextualized representations h_x and h_y, respectively. Following prior work (Jalili Sabet et al., 2020; Dou and Neubig, 2021; Wang et al., 2022), we next obtain the token similarity matrix S_{xy}:

$$S_{xy} = h_x h_y^T$$

We row- and column-normalize the similarity matrix using softmax to obtain $\hat{S}_{xy}$ and $\hat{S}_{yx}$—capturing the similarity from x to y and y to x. Finally, we compute the alignment matrix A by intersecting the two similarity matrices:

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

where c is the alignment threshold and $A_{ij} = 1$ indicates that two tokens are aligned. As the MT encoder operates on the level of sub-word tokens, we consider two words to be aligned if any of their sub-word tokens are aligned, in line with the prior WA work (Jalili Sabet et al., 2020; Dou and Neubig, 2021; Wang et al., 2022).

Fine-Tuning for Word Alignment. Additionally, we explore fine-tuning TransAlign on a word alignment-specific objective to further improve performance. Different from related work—that employed full fine-tuning (Dou and Neubig, 2021; Wang et al., 2022) or adapter-based fine-tuning (Wang et al., 2022)—we opt for LoRA (Hu et al., 2022) as it does not increase model depth while maintaining parameter efficiency. We resort to the following loss function for WA fine-tuning:

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

where $\hat{A}$ refers to the alignments extracted at the current training step and n (or m) is the number of tokens in sentence x (or y) (Dou and Neubig, 2021; Wang et al., 2022).

3 Experiments

With label projection as the key remaining application of word aligners, we first evaluate TransAlign on translation-based XLT for token classification. We then additionally benchmark TransAlign intrinsically on word alignment itself.

3.1 Experimental Setup

TransAlign. For both extrinsic and intrinsic evaluation, we use the encoder of the distilled 600M parameter version of NLLB (Team et al., 2022) as the backbone of TransAlign. We extract alignments after the last (i.e., 12th) layer using an alignment threshold of $c = 0.001$.

Extrinsic Evaluation. We benchmark the downstream capabilities of the fine-tuned TransAlign on translation-based XLT for token classification.

We resort to `T-Test`—since it is shown to outperform `T-Train` (Le et al., 2024; Ebing and Glavaš, 2025)—where an MT model is used to translate the original target language instances into the (noisy) source language before inference. Afterward, the predictions are mapped back to the original target language via word alignment. Having obtained the alignments, we follow the span-based label projection algorithm of Ebing and Glavaš (2025).[1]

Evaluation Tasks. We evaluate for 28 diverse languages on two established token classification tasks: named entity recognition (NER) and slot labeling (SL). For NER, we use MasakhaNER2.0 (Masakha) (Adelani et al., 2022) encompassing low-resource languages from Sub-Saharan Africa. For SL, the evaluation dataset is xSID (van der Goot et al., 2021), covering mid- to high-resource languages and dialects. In all experiments, we use English as the source language.

Label Projection Baselines. Our baselines comprise two popular WAs based on multilingual encoders (i) AwsmAlign (Dou and Neubig, 2021), based on multilingual BERT, and (ii) AccAlign (Wang et al., 2022), based on the multilingual sentence encoder LaBSE. Moreover, we include Codec (Le et al., 2024)—a state-of-the-art non-WA-based label projection method that identifies labeled spans in the translated sentence post-translation by means of constrained decoding.

Downstream Fine-Tuning. We evaluate XLM-R Large (Conneau et al., 2020) and DeBERTaV3 Large (He et al., 2023) as our downstream LMs. We train the models on the original English data and run experiments with 3 random seeds. We report the mean F_1 score and standard deviation.

Intrinsic Evaluation. We evaluate TransAlign on 8 language pairs: en-cz/de/fr/hi/ja/ro/sv/zh and compare it against the same WA-baselines. All WAs are evaluated in their vanilla (non-fine-tuned) variant. We report AER for each language pair. We provide full details of the intrinsic evaluation in the Appendix B.

3.2 Main Results

Extrinsic Evaluation. Table 1 outlines the `T-Test` results for the fine-tuned WAs and Codec. We demonstrate that all `T-Test` strategies exceed zero-shot XLT substantially reaching an improvement of up to 13.4% on average (with TransAlign and XLM-R). Comparing TransAlign against the other WA baselines, we find it to clearly outperform AwsmAlign and AccAlign by 5.5% and 1.5% on average.[2] Not only does TransAlign outperform popular WAs in translation-based XLT for token classification, but it also improves over the competitive non-WA-based label projection method Codec by 2% on average. This finding is noteworthy as TransAlign is a *fair* baseline for Codec: both approaches use a fine-tuned NLLB model of the same size for label projection. However, TransAlign is computationally more efficient as it only uses the encoder of NLLB and thus avoids the costly constrained decoding of Codec (Le et al., 2024).

		Masakha	xSID	Avg
ZS	X	$52.9_{\pm 1.8}$	$76.5_{\pm 1.4}$	$64.7_{\pm 1.7}$
Translate-Test: non-WA				
Codec	X	$72.0_{\pm 0.5}$	$80.1_{\pm 0.3}$	$76.1_{\pm 0.4}$
Codec	D	$72.4_{\pm 0.4}$	$80.2_{\pm 0.4}$	$76.3_{\pm 0.4}$
Translate-Test: WA				
AwsmAlign	X	$68.4_{\pm 0.4}$	$78.8_{\pm 0.3}$	$73.6_{\pm 0.4}$
AwsmAlign	D	$68.8_{\pm 0.4}$	$78.7_{\pm 0.4}$	$73.8_{\pm 0.4}$
AccAlign	X	$72.3_{\pm 0.4}$	$80.9_{\pm 0.3}$	$76.6_{\pm 0.4}$
AccAlign	D	$72.7_{\pm 0.4}$	$80.8_{\pm 0.4}$	$76.8_{\pm 0.4}$
TransAlign	X	$73.9_{\pm 0.4}$	$\mathbf{82.2}_{\pm 0.4}$	$78.1_{\pm 0.4}$
TransAlign	D	$\mathbf{74.3}_{\pm 0.4}$	$\mathbf{82.2}_{\pm 0.4}$	$\mathbf{78.3}_{\pm 0.4}$

Table 1: Main results for translation-based XLT for token classification. Results with XLM-R (X) and DeBERTa (D). We report mean F1.

	en-zh	en-cs	en-fr	en-de	en-hi	en-ja	en-ro	en-sv
All Words								
AwsmAlign	18.2	12.3	6.3	18.6	42.9	46.2	28.9	9.9
AccAlign	**16.2**	9.3	**5.2**	**16.4**	30.4	43.3	20.8	**7.3**
TransAlign	18.8	**8.9**	6.8	17.7	**29.4**	**43.2**	**20.6**	7.8
w/o Stopwords								
AwsmAlign	12.5	10.6	5.3	14.2	35.6	**35.3**	22.0	9.2
AccAlign	10.7	6.8	4.3	**11.6**	24.9	37.5	16.1	5.8
TransAlign	**10.6**	**6.3**	**4.0**	11.8	**23.4**	36.5	**15.2**	**5.1**

Table 2: Main results for word alignment evaluation. Word alignment models are evaluated in their vanilla (non-fine-tuned) variant. We report the AER considering all words and without considering stopwords.

Intrinsic Evaluation. We present the results for intrinsic evaluation in Table 2. Considering all words in the source and target sentence equally, we

[1] The algorithm projects labels across spans and not individual tokens and can compensate for some word alignment errors. For details, we refer the reader to the original work.

[2] Since TransAlign covers substantially more languages than AccAlign, we provide additional experiments demonstrating that the improved performance does not stem from broader language coverage (see Appendix E).

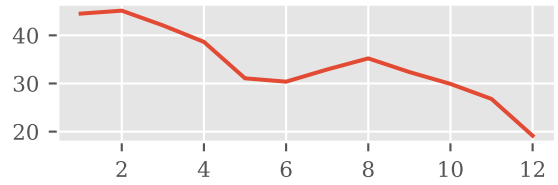

Figure 1: Word alignment performance across layers of vanilla TransAlign. We present the average AER over all 8 language pairs.

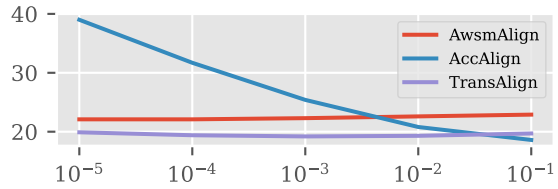

Figure 2: Word alignment performance for different thresholds of c. We evaluate vanilla WAs and present the average AER over all 8 language pairs.

find that TransAlign produces the best results for 4 out of 8 language pairs (AccAlign reaches the best performance on the remaining ones). While TransAlign and AccAlign perform similarly on alignment itself, our TransAlign exhibited stronger downstream XLT performance (Table 1). For example, in intrinsic evaluation, AccAlign outperforms TransAlign for Chinese and German by 2.6% and 1.3%, respectively. In contrast, for `T-Test` on xSID (see App. G), the trend turns around: TransAlign outperforms AccAlign for both Chinese (0.7%) and German (2.2%).

These results point to a mismatch between the standard word alignment evaluation that treats each word in the input as equally important and label projection for translation-based XLT that requires correct alignments on a subset of the input sentence. Commonly, labeled spans in downstream evaluation span words that carry meaning (e.g., named entities). We thus additionally report the alignment results by excluding stopwords—words with little semantic meaning—from the evaluation. Results presented in Table 2 (bottom half) support our hypothesis: not accounting for the (accuracy of) stopword alignment, TransAlign outperforms both baselines consistently: this means it produces more accurate alignments between content words, which explains why it yields downstream XLT gains.

3.3 Analysis

Performance per Layer. The layer from which we extract the alignments can have a substantial impact on performance (Jalili Sabet et al., 2020; Dou and Neubig, 2021; Wang et al., 2022). Figure 1 shows the average AER performance for all layers of vanilla TransAlign: using the last layer of TransAlign substantially outperforms using any other layer.

Alignment Threshold. The threshold parameter c decides whether two tokens are considered to be aligned. We ablate the choice of c for all WAs in their vanilla variant (see Figure 2). While AwsmAlign and TransAlign are robust to the threshold value, we find AccAlign's performance to severely vary with the value of c.

	Masakha	xSID	Avg
Non-Fine-Tuned WAs			
AwsmAlign	$66.2_{\pm 0.3}$	$74.1_{\pm 0.3}$	$70.2_{\pm 0.3}$
AccAlign	$71.2_{\pm 0.4}$	$80.0_{\pm 0.4}$	$75.6_{\pm 0.4}$
TransAlign	$73.5_{\pm 0.4}$	$81.8_{\pm 0.4}$	$77.7_{\pm 0.4}$
Fine-Tuned WAs			
AwsmAlign	$68.8_{\pm 0.4}$	$78.7_{\pm 0.4}$	$73.8_{\pm 0.4}$
AccAlign	$72.7_{\pm 0.4}$	$80.8_{\pm 0.4}$	$76.8_{\pm 0.4}$
TransAlign	$\mathbf{74.3}_{\pm 0.4}$	$\mathbf{82.2}_{\pm 0.4}$	$\mathbf{78.3}_{\pm 0.4}$

Table 3: Impact of WA fine-tuning on translation-based XLT for token classification. Results with DeBERTa.

Impact of WA Fine-Tuning. We obtained the best results for our WA-fine-tuned TransAlign (Table 1). We next assess the contribution of word alignment fine-tuning for all WAs on downstream MT-based XLT performance (see Table 3). We find that fine-tuning improves the XLT results for all WAs, but the gains are more pronounced for WAs with weaker initial performance: AwsmAlign improves by 3.6% compared to 0.6% for TransAlign. We also note that using a stronger WA model is more beneficial than fine-tuning: vanilla TransAlign outperforms the WA-fine-tuned AccAlign by 0.7%.

4 Conclusion

In this work, we proposed TransAlign, a new word aligner (WA) that leverages the encoder of NLLB, a massively multilingual encoder-decoder MT model. Our extrinsic evaluation on translation-based XLT for token classification on two established benchmarks covering 28 languages, shows that TransAlign outperforms popular existing WAs as well as state-of-the-art non-WA-based label projection methods. Furthermore, our intrinsic word alignment evaluation reveals that, TransAlign aligns content words (rather than functional words) in particular better than existing WAs, which then reflects in downstream XLT gains.

5 Limitations

We focused on choosing well-established and representative tasks for token classification. However, in NLP, multilingual evaluation benchmarks are often created by translating the data from an existing high-resource language followed by post-editing. This applies to xSID and some languages of Masakha. As a result, the newly introduced languages might contain translation artifacts referred to as *translationese*. Prior work (Artetxe et al., 2020, 2023) stated that translation-based XLT strategies might lead to exploitation of translationese, slightly overestimating performance.

Our intrinsic evaluation points to a potential mismatch between the word alignment task and the extrinsic evaluation on translation-based XLT for token classification. Our results suggest that the mismatch stems from the discrepancy of treating all words equally (intrinsic evaluation) against focusing on a specific subset of words (extrinsic evaluation). While we hypothesize as to why MT models perform worse in aligning words with little semantic meaning than sentence encoders, further work is needed to test our hypothesis.

References

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumim, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Niraj Aswani and Robert Gaizauskas. 2005. Aligning words in English-Hindi parallel corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 115–118, Ann Arbor, Michigan. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 175–182, Barcelona, Spain.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *Preprint*, arXiv:1805.10190.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 2112–2128, Online. Association for Computational Linguistics.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised POS tagging with bilingual projections. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 634–639, Sofia, Bulgaria. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Benedikt Ebing and Goran Glavaš. 2024. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5325–5344, Mexico City, Mexico. Association for Computational Linguistics.

Benedikt Ebing and Goran Glavaš. 2025. The devil is in the word alignment details: On translation-based cross-lingual transfer for token classification tasks. *Preprint*, arXiv:2505.10507.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Michael Elhadad. 2010. Book review: Natural language processing with python by steven bird, ewan Klein, and edward loper. *Computational Linguistics*, 36(4).

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831, Online. Association for Computational Linguistics.

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

Javier Ferrando and Marta R. Costa-jussà. 2021. Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 434–443, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2023. T-projection: High quality annotation projection for sequence labeling tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15203–15217, Singapore. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Maria Holmqvist and Lars Ahrenberg. 2011. A gold standard for English-Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 106–113, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

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.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2295–2301. AAAI Press.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic alignment of Czech and English deep syntactic dependency trees. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, pages 103–113, Hamburg, Germany. European Association for Machine Translation.

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

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

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

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

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual label projection for cross-lingual structured prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2023. Transfer-free data-efficient multilingual slot labeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6041–6055, Singapore. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. 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 910–920, Hong Kong, China. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In

Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142–147.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

David Vilar, Maja Popovic, and Hermann Ney. 2006. AER: do we need to "improve" our alignments? In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*, Kyoto, Japan.

Weikang Wang, Guanhua Chen, Hanqing Wang, Yue Han, and Yun Chen. 2022. Multilingual sentence transformer as a multilingual word aligner. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2952–2963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

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.

A Experimental Details: Extrinsic Evaluation

Machine Translation. For translation, we utilize the state-of-the-art massively multilingual NLLB model with 3.3B parameters (Team et al., 2022). Following prior work (Artetxe et al., 2023; Ebing and Glavaš, 2024; Ebing and Glavaš, 2025), we decode using beam search with a beam size of 5. For Masakha (Adelani et al., 2022) and xSID (van der Goot et al., 2021), we concatenated the pre-tokenized input on whitespace before translation. We deviate from this for the Chinese data in xSID, where we merge Chinese tokens without whitespace. Additionally, the dialect *South Tyrol* (de-st) in xSID is not supported by NLLB. We translate the dialect pretending it to be German (i.e., using the German language code) as it is closely related to the latter. We accessed all datasets through the Hugging Face library and ensured compliance with the licenses. All translations were run on a single A100 with 40GB VRAM.

Word Aligners. We will publicly release our word alignment code (Apache 2.0 license) and the model checkpoints for the fine-tuned TransAlign (CC-BY-NC 4.0 license). Next to TransAlign, we re-implemented two popular word aligners as our baselines: AwsmAlign (Dou and Neubig, 2021) and AccAlign (Wang et al., 2022). We chose the code repository of SimAlign (Jalili Sabet et al., 2020) as the starting point for our implementation. We accessed the code through their repository: (https://github.com/cisnlp/simalign). Following Dou and Neubig (2021), we extracted alignments for AwsmAlign after the 8th layer using an alignment threshold of $c = 0.001$. For AccAlign, we use the 6th layer and an alignment threshold of $c = 0.1$ (Wang et al., 2022). We comply with the licenses of AwsmAlign (BSD 3-Clause) and SimAlign (MIT). We could not find licensing information for AccAlign.

Codec. Codec (Le et al., 2024) is a label projection method that leverages constrained decoding as part of a two-step translation procedure. In the first step, the source sentence is translated into the target language (e.g., from English: "This is New York" to German: "Das ist New York"). Then, in the second step, tags are inserted around the labeled spans in the source sentence (English: "This is [New York]"). The marked sentence is fed again as input to the MT model: during decoding, the MT model is now constrained to generate only the tokens from the translation obtained in the first step ("Das", "ist", "New", "York") or a tag ("[", "]"). We chose Codec as a representative method for non-WA-based label projection: Ebing and Glavaš (2025) suggest that Codec performs on par or better than comparable non-WA-based label projection methods (Chen et al., 2023; García-Ferrero et al., 2023; Parekh et al., 2024). To project the labels for `T-Test`, we used the publicly available code repository of Codec: https://github.com/duonglm38/Codec. While an implementation for Masakha is already provided, we extended their implementation to handle label projection for xSID. We adhered to the hyperparameters in their repository and followed the existing implementation closely. The constrained decoding (i.e., inserting the tags post-translation) requires a fine-tuned NLLB that is able to preserve/insert

tags. Therefore, we follow Le et al. (2024) using the fine-tuned 600M parameter version of NLLB released by Chen et al. (2023). We could not find licensing information for Codec.

Label Projection. We follow the span-based label projection procedure used by (Ebing and Glavaš, 2025). The algorithm projects labels across spans and not individual tokens and can compensate for some word alignment errors. For details, we refer the reader to the original work. Unlike their work, we do not apply filtering heuristics for `T-Test`.

Word Aligner Fine-Tuning. For fine-tuning, we apply LoRA to the feed-forward sublayer of each encoder layer. We train each WA for 20 epochs using a learning rate of $1e^{-4}$. The rank is set to 8 and alpha to 32. We apply LoRA dropout with 0.01. For WA training, we utilize the labeled data from the intrinsic evaluation (see Table 7).

Downstream Fine-Tuning. We train both tasks (NER and SL) for 10 epochs using an effective batch size of 32. In case we can not fit the desired batch size, we utilize gradient accumulation. The learning rate is set to $1e^{-5}$ with a weight decay of 0.01. We implement a linear schedule of 10% warm-up and employ mixed precision. We evaluate models at the last checkpoint of training. We use the seqeval F1 implementation accessed through the Hugging Face library. Further, we access our downstream models—XLM-RoBERTa Large and DeBERTaV3 Large—through the Hugging Face library. All downstream training and evaluation runs were completed on a single V100 with 32GB VRAM. We estimate the GPU time to be 2000 hours across all translations and downstream fine-tunings.

Datasets.

MasakhaNER2.0. Our experiments cover 18 out of 20 languages that are supported by NLLB. Note that Google Translate (GT) does not support all 18 languages. Following, we mark the 11 languages that are supported by GT with an additional asterisk: Bambara (bam)*, Ewé (ewe)*, Fon (fon), Hausa (hau)*, Igbo (ibo)*, Kinyarwanda (kin)*, Luganda (lug), Luo (luo), Mossi (most), Chichewa (nya), chiShona (sna)*, Kiswahili (saw)*, Setswana (tsn), Akan/Twi (twi)*, Wolof (wol), isiXhosa (xho)*, Yorùrbá (yor)*, and isiZulu (zul)*. As source data, we use the English training (14k instances) and validation portions (3250 instances) of CoNLL (Tjong Kim Sang and De Meulder, 2003).

xSID. We evaluate 10 languages all covered by NLLB and GT: Arabic (ar), Danish (da), German (de), South-Tyrolean (de-st), Indonesian (id), Italian (it), Kazakh (kk), Dutch (nl), Turkish (tr), and Chinese (zh). Following Razumovskaia et al. (2023), we excluded Japanese from the evaluation because it only has half of the validation and test instances and spans only a fraction of entities compared to the other languages. Moreover, we exclude Serbian as the evaluation data is written in the Latin script whereas NLLB was only trained in the Cyrillic script. xSID is an evaluation-only dataset. Therefore, we follow van der Goot et al. (2021) and use their publicly released English data for training and validation. The instances are sourced from the Snips (Coucke et al., 2018) and Facebook (Schuster et al., 2019) SL datasets. We deduplicate the training instances, ending up with over 36k training and 300 validation examples.

B Experimental Details: Intrinsic Evaluation

Word Alignment Baselines. We use the same WA models as for the extrinsic evaluation—AwsmAlign and AccAlign (see App. A). All WAs are evaluated in their non-fine-tuned variant.

Languages. We evaluate the following 8 language pairs: English-Chinese (en-zh), English-Czech (en-cz), English-French (en-fr), English-German (en-de), English-Hindi (en-hi), English-Japanese (en-ja), English-Romanian (en-ro) and English-Swedish (en-sv). We provide details on the used datasets in Table 7.

Stopword Filtering. For the results in Table 2, we applied stopword filtering prior to AER computation. We identified stopwords from the English source sentences using the stopword list provided by NLTK (Elhadad, 2010) and removed corresponding target language words accordingly. The NLTK source code is published under the Apache 2.0 license. We comply with their license.

C Further Analysis: Robustness of Fine-Tuning

For the application of a fine-tuned WA model, only a single seed of a fine-tuned model will eventually be used. Therefore, we ablate the variance of the random seed chosen for fine-tuning. We fine-tune AwsmAlign, AccAlign, and TransAlign on three distinct random seeds and evaluate them on

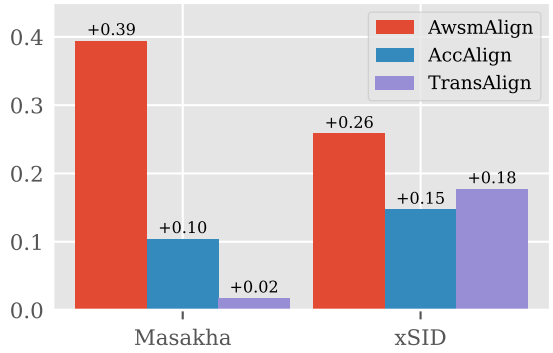

Figure 3: Variance of WA model fine-tuning with three distinct random seeds evaluated on translation-based XLT. Results with DeBERTa.

translation-based XLT. The resulting variance is depicted in Figure 3. We observe little impact by the choice of the random seed for TransAlign: for xSID the variance is comparable to that of AwsmAlign and AccAlign, while for Masakha, it is substantially lower.

D Further Analysis: MT Model

In translation-based XLT for token classification, it is pragmatic to use the encoder of the MT model for word alignment since (i) only a single model is required for the label projection pipeline (i.e., translation and label projection) and (ii) the language coverage of target languages is ensured for both steps. However, open access to the encoder of the MT model is required. With closed commercial MT models being considered to produce superior translation quality, we explore whether the gains obtained by TransAlign are orthogonal to the MT model. Our results in Table 4 suggest that TransAlign does not depend on its *own* translations. The performance improvements obtained by label projection with TransAlign are orthogonal to gains obtained by higher translation quality.

		Masakha	xSID	Avg
AwsmAlign	NLLB	$69.2_{\pm 0.4}$	$78.7_{\pm 0.4}$	$74.0_{\pm 0.4}$
AccAlign	NLLB	$73.7_{\pm 0.4}$	$80.8_{\pm 0.4}$	$77.3_{\pm 0.4}$
TransAlign	NLLB	$75.1_{\pm 0.5}$	$82.2_{\pm 0.4}$	$78.7_{\pm 0.5}$
AwsmAlign	GT	$70.6_{\pm 0.3}$	$80.1_{\pm 0.4}$	$75.3_{\pm 0.4}$
AccAlign	GT	$75.2_{\pm 0.4}$	$82.1_{\pm 0.4}$	$78.6_{\pm 0.4}$
TransAlign	GT	$76.4_{\pm 0.5}$	$83.6_{\pm 0.4}$	$80.0_{\pm 0.4}$

Table 4: Results for translation-based XLT for token-level tasks with translations obtained from different MT models—Google Translation (GT) and NLLB (NLLB). Results with DeBERTa.

E Further Analysis: Language Coverage

NLLB has seen substantially more languages in pretraining than LaBSE (200 vs. 109 languages). To ensure that performance improvements obtained by TransAlign do not simply stem from broader language coverage, we evaluate TransAlign and AccAlign on a subset of languages seen in the pretraining of both models. We observe that TransAlign still outperforms AccAlign even on a subset of languages seen by both models (see Table 5).

	Masakha	xSID	Avg
AccAlign	$74.1_{\pm 0.5}$	$83.2_{\pm 0.4}$	$78.7_{\pm 0.4}$
TransAlign	$75.4_{\pm 0.5}$	$84.7_{\pm 0.4}$	$80.0_{\pm 0.4}$

Table 5: Results for translation-based XLT for token-level tasks only evaluating languages seen in the pretraining of both WAs. Results with DeBERTa.

F Further Analysis: NLLB Model Size

NLLB is released in different model sizes ranging from 600M up to 54B parameters. Table 6 compares the fine-tuned TransAlign in two different model sizes. We evaluate the 600M (distilled) and 3.3B parameter models on translation-based XLT for token classification. Our results reveal that the larger model does not provide any advantage. Hence, we used the 600M parameter model for our main results.

		Masakha	xSID	Avg
TransAlign	600M	$74.3_{\pm 0.4}$	$82.2_{\pm 0.4}$	$78.3_{\pm 0.4}$
TransAlign	3.3B	$74.5_{\pm 0.4}$	$81.4_{\pm 0.4}$	$78.0_{\pm 0.4}$

Table 6: Results for translation-based XLT for token-level tasks with different sizes of NLLB as WA. Results with DeBERTa.

Lang	Source	Link	#Sents
en-zh	(Liu and Sun, 2015)	https://nlp.csai.tsinghua.edu.cn/ ly/systems/TsinghuaAligner/TsinghuaAligner.html	450
en-cs	(Mareček et al., 2008)	https://ufal.mff.cuni.cz/czech-english-manual-word-alignment	2400
en-fr	(Mihalcea and Pedersen, 2003)	https://web.eecs.umich.edu/ mihalcea/wpt/	447
en-de	(Vilar et al., 2006)	https://www-i6.informatik.rwth-aachen.de/goldAlignment/	508
en-hi	(Aswani and Gaizauskas, 2005)	https://web.eecs.umich.edu/ mihalcea/wpt05/	90
en-ja	(Neubig, 2011)	https://www.phontron.com/kftt/	582
en-ro	(Mihalcea and Pedersen, 2003)	https://web.eecs.umich.edu/ mihalcea/wpt05/	248
en-sv	(Holmqvist and Ahrenberg, 2011)	https://www.ida.liu.se/divisions/hcs/nlplab/resources/ges/	192

Table 7: Datasets used for intrinsic evaluation and fine-tuning of WAs. For the fine-tuning, we held out 100 randomly selected instances of the en-cs dataset as validation portion.

G Detailed Results: Main Results

		bam	ewe	fon	hau	ibo	kin	lug	luo	mos	nya	sna	swa	tsn	twi	wol	xho	yor	zul	Avg
ZS	X	43.4	72.8	61.0	73.5	49.9	46.3	64.9	55.0	56.1	51.1	34.4	88.1	51.5	49.5	56.2	22.2	35.1	41.5	52.9
Translate-Test: non-WA																				
Codec	X	54.5	78.8	67.4	72.9	72.8	77.6	83.6	72.8	49.4	78.1	79.3	82.2	79.2	72.5	67.3	72.5	58.4	77.1	72.0
Codec	D	54.3	79.1	68.0	73.3	73.9	78.2	83.5	74.2	48.8	79.0	79.8	82.9	79.3	73.1	67.8	72.6	58.0	77.0	72.4
Translate-Test: WA																				
AwsmAlign	X	51.4	78.7	61.3	70.9	75.4	66.8	82.7	72.2	47.6	77.5	71.8	81.5	79.5	70.8	62.1	56.0	61.7	63.8	68.4
AwsmAlign	D	51.1	78.8	62.1	71.4	77.0	67.6	82.5	73.6	47.6	77.9	72.3	82.1	79.8	71.7	62.5	56.2	61.2	63.8	68.8
AccAlign	X	54.4	79.9	69.7	74.7	75.2	70.8	84.4	72.6	53.1	78.6	81.7	83.0	80.0	71.2	64.9	73.2	55.4	78.7	72.3
AccAlign	D	53.8	79.9	70.1	75.2	76.7	71.5	84.2	74.1	53.2	79.1	82.3	83.6	80.4	72.0	65.4	73.3	55.3	78.8	72.7
TransAlign	X	56.8	80.8	72.8	74.9	75.8	71.0	84.8	74.7	54.0	78.8	82.2	82.3	82.2	75.1	68.6	73.8	62.8	79.0	73.9
TransAlign	D	56.6	80.8	73.3	75.4	77.3	71.7	84.6	76.3	53.7	79.2	82.8	82.9	82.6	75.8	69.3	74.0	62.5	79.1	74.3

Table 8: Detailed main results for translation-based XLT on Masakha. Results with XLM-R (X) and DeBERTa (D).

		ar	da	de	de-st	id	it	kk	nl	tr	zh	Avg
ZS	X	71.5	85.6	80.8	43.9	86.8	88.2	80.8	88.8	81.5	57.4	76.5
Translate-Test: non-WA												
Codec	X	79.0	81.9	86.1	60.4	84.8	88.4	83.0	86.5	83.6	67.0	80.1
Codec	D	79.9	81.8	85.5	58.8	85.8	89.0	83.2	86.0	84.2	67.5	80.2
Translate-Test: WA												
AwsmAlign	X	79.1	76.2	85.2	60.2	79.1	87.9	75.3	87.3	78.1	80.1	78.8
AwsmAlign	D	79.3	75.9	84.4	58.4	79.9	88.6	75.1	86.5	78.9	80.0	78.7
AccAlign	X	80.2	75.8	85.2	61.0	84.1	88.2	82.6	86.6	82.4	82.5	80.9
AccAlign	D	80.7	75.5	84.5	59.3	85.0	88.9	82.7	86.0	83.3	82.4	80.8
TransAlign	X	81.0	81.0	87.4	61.0	87.0	88.5	82.4	87.8	83.2	83.2	82.2
TransAlign	D	81.4	80.7	86.7	59.3	87.9	89.2	82.4	86.9	84.1	83.1	82.2

Table 9: Detailed main results for translation-based XLT on xSID. Results with XLM-R (X) and DeBERTa (D).

H Detailed Results: Impact of Fine-Tuning

		bam	ewe	fon	hau	ibo	kin	lug	luo	mos	nya	sna	swa	tsn	twi	wol	xho	yor	zul	Avg
Non-Fine-Tuned WAs																				
AwsmAlign	X	46.0	76.9	57.9	70.1	75.5	64.9	83.0	71.8	43.5	77.9	63.3	79.8	80.6	70.9	53.1	50.0	58.1	60.3	65.8
AwsmAlign	D	46.0	77.0	58.6	70.6	76.9	65.6	82.8	73.2	43.6	78.4	63.7	80.5	81.2	71.7	53.5	50.0	57.7	60.2	66.2
AccAlign	X	54.7	79.1	68.2	74.1	72.7	69.7	83.6	70.7	49.5	77.5	80.5	81.3	81.3	71.9	63.2	70.9	48.3	76.9	70.8
AccAlign	D	54.1	79.1	68.6	74.6	74.0	70.2	83.3	72.0	49.3	78.2	81.0	82.0	81.6	73.0	63.6	71.1	48.2	77.0	71.2
TransAlign	X	55.6	80.1	70.5	74.6	75.0	70.4	84.8	73.6	52.4	77.8	81.5	82.2	82.2	75.2	67.2	73.4	60.1	78.6	73.1
TransAlign	D	55.4	80.1	71.2	75.1	76.6	71.0	84.6	75.0	52.4	78.5	82.1	82.8	82.5	75.9	68.0	73.5	59.9	78.8	73.5
Fine-Tuned WAs																				
AwsmAlign	X	51.4	78.7	61.3	70.9	75.4	66.8	82.7	72.2	47.6	77.5	71.8	81.5	79.5	70.8	62.1	56.0	61.7	63.8	68.4
AwsmAlign	D	51.1	78.8	62.1	71.4	77.0	67.6	82.5	73.6	47.6	77.9	72.3	82.1	79.8	71.7	62.5	56.2	61.2	63.8	68.8
AccAlign	X	54.4	79.9	69.7	74.7	75.2	70.8	84.4	72.6	53.1	78.6	81.7	83.0	80.0	71.2	64.9	73.2	55.4	78.7	72.3
AccAlign	D	53.8	79.9	70.1	75.2	76.7	71.5	84.2	74.1	53.2	79.1	82.3	83.6	80.4	72.0	65.4	73.3	55.3	78.8	72.7
TransAlign	X	56.8	80.8	72.8	74.9	75.8	71.0	84.8	74.7	54.0	78.8	82.2	82.3	82.2	75.1	68.6	73.8	62.8	79.0	73.9
TransAlign	D	56.6	80.8	73.3	75.4	77.3	71.7	84.6	76.3	53.7	79.2	82.8	82.9	82.6	75.8	69.3	74.0	62.5	79.1	74.3

Table 10: Impact of WA fine-tuning on translation-based XLT on Masakha. Results with XLM-R (X) and DeBERTa (D).

		ar	da	de	de-st	id	it	kk	nl	tr	zh	Avg
Non-Fine-Tuned WAs												
AwsmAlign	X	74.2	75.8	84.6	58.9	76.0	85.3	59.3	85.7	69.2	73.6	74.1
AwsmAlign	D	74.8	75.5	83.8	56.9	76.4	86.0	59.7	85.2	69.7	73.3	74.1
AccAlign	X	78.8	75.4	84.8	59.5	82.0	86.4	81.8	86.6	82.7	82.2	80.0
AccAlign	D	79.2	75.2	84.1	57.9	83.0	87.2	81.8	85.9	83.6	82.1	80.0
TransAlign	X	80.0	81.0	87.3	61.0	86.5	87.8	81.8	87.3	83.8	82.6	81.9
TransAlign	D	80.3	80.7	86.6	59.4	87.5	88.5	81.8	86.6	84.6	82.6	81.8
Fine-Tuned WAs												
AwsmAlign	X	79.1	76.2	85.2	60.2	79.1	87.9	75.3	87.3	78.1	80.1	78.8
AwsmAlign	D	79.3	75.9	84.4	58.4	79.9	88.6	75.1	86.5	78.9	80.0	78.7
AccAlign	X	80.2	75.8	85.2	61.0	84.1	88.2	82.6	86.6	82.4	82.5	80.9
AccAlign	D	80.7	75.5	84.5	59.3	85.0	88.9	82.7	86.0	83.3	82.4	80.8
TransAlign	X	81.0	81.0	87.4	61.0	87.0	88.5	82.4	87.8	83.2	83.2	82.2
TransAlign	D	81.4	80.7	86.7	59.3	87.9	89.2	82.4	86.9	84.1	83.1	82.2

Table 11: Impact of WA fine-tuning on translation-based XLT on xSID. Results with XLM-R (X) and DeBERTa (D).

I Detailed Results: MT Model

		bam	ewe	hau	ibo	kin	sna	swa	twi	xho	yor	zul	Avg
AwsmAlign	NLLB	51.1	78.8	71.4	77.0	67.6	72.3	82.1	71.7	56.2	61.2	63.8	69.2
AccAlign	NLLB	53.8	79.9	75.2	76.7	71.5	82.3	83.6	72.0	73.3	55.3	78.8	73.7
TransAlign	NLLB	56.6	80.8	75.4	77.3	71.7	82.8	82.9	75.8	74.0	62.5	79.1	75.1
AwsmAlign	GT	55.4	78.9	71.9	79.4	68.1	75.2	84.1	73.5	59.0	65.1	66.1	70.6
AccAlign	GT	59.6	79.3	74.2	79.3	72.4	84.7	86.0	73.5	75.2	61.6	81.2	75.2
TransAlign	GT	61.5	79.9	74.2	80.4	72.6	84.8	86.11	77.13	75.73	66.73	81.3	76.4

Table 12: Detailed results for translation-based XLT on Masakha with translations obtained from different MT models—Google Translation (GT) and NLLB (NLLB). Results with DeBERTa.

		ar	da	de	de-st	id	it	kk	nl	tr	zh	Avg
AwsmAlign	NLLB	79.3	75.9	84.4	58.4	79.9	88.6	75.1	86.5	78.9	80.0	78.7
AccAlign	NLLB	80.7	75.5	84.5	59.3	85.0	88.9	82.7	86.0	83.3	82.4	80.8
TransAlign	NLLB	81.4	80.7	86.7	59.3	87.9	89.2	82.4	86.9	84.1	83.1	82.2
AwsmAlign	GT	81.3	76.0	85.6	58.9	79.7	90.2	76.3	87.8	82.0	83.2	80.1
AccAlign	GT	81.7	76.6	85.3	58.8	85.3	90.1	85.1	87.1	84.4	86.4	82.1
TransAlign	GT	82.6	81.4	87.6	58.8	87.1	91.9	84.6	89.0	86.2	86.7	83.6

Table 13: Detailed results for translation-based XLT on xSID with translations obtained from different MT models—Google Translation (GT) and NLLB (NLLB). Results with DeBERTa.

J Detailed Results: Language Coverage

	hau	ibo	kin	nya	sna	swa	wol	xho	yor	zul	Avg
AccAlign	75.2	76.7	71.5	79.1	82.3	83.6	65.4	73.3	55.3	78.8	74.1
TransAlign	75.4	77.3	71.7	79.2	82.8	82.9	69.3	74.0	62.5	79.1	75.4

Table 14: Detailed results for translation-based XLT on Masakha only evaluating languages seen in the pretraining of both WAs. Results with DeBERTa.

	ar	da	de	id	it	kk	nl	tr	zh	Avg
AccAlign	80.7	75.5	84.5	85.0	88.9	82.7	86.0	83.3	82.4	83.2
TransAlign	81.4	80.7	86.7	87.9	89.2	82.4	86.9	84.1	83.1	84.7

Table 15: Detailed results for translation-based XLT on xSID only evaluating languages seen in the pretraining of both WAs. Results with DeBERTa.

K Detailed Results: NLLB Model Size

		bam	ewe	fon	hau	ibo	kin	lug	luo	mos	nya	sna	swa	tsn	twi	wol	xho	yor	zul	Avg
TransAlign	600M	56.6	80.8	73.3	75.4	77.3	71.7	84.6	76.3	53.7	79.2	82.8	82.9	82.6	75.8	69.3	74.0	62.5	79.1	74.3
TransAlign	3.3B	57.1	80.7	74.0	75.2	77.3	71.8	84.6	76.3	54.1	79.6	82.7	83.3	82.5	75.8	69.5	74.0	62.8	79.2	74.5

Table 16: Detailed results for translation-based XLT on Masakha with different sizes of NLLB as WA. Results with DeBERTa.

		ar	da	de	de-st	id	it	kk	nl	tr	zh	Avg
TransAlign	600M	81.4	80.7	86.7	59.3	87.9	89.2	82.4	86.9	84.1	83.1	82.2
TransAlign	3.3B	80.8	76.0	86.2	59.3	82.5	89.4	83.5	86.8	86.3	83.7	81.4

Table 17: Detailed results for translation-based XLT on xSID with different sizes of NLLB as WA. Results with DeBERTa.