Exploring the Relationship between Alignment and Cross-lingual Transfer in Multilingual Transformers

Félix Gaschi [1,2], **Patricio Cerda** [1], **Parisa Rastin** [2], **Yannick Toussaint** [2]
[1]Posos, [2]LORIA
{felix.gaschi,parisa.rastin,yannick.toussaint}@loria.fr
patricio@posos.fr

Abstract

Without any explicit cross-lingual training data, multilingual language models can achieve cross-lingual transfer. One common way to improve this transfer is to perform realignment steps before fine-tuning, i.e., to train the model to build similar representations for pairs of words from translated sentences. But such realignment methods were found to not always improve results across languages and tasks, which raises the question of whether aligned representations are truly beneficial for cross-lingual transfer. We provide evidence that alignment is actually significantly correlated with cross-lingual transfer across languages, models and random seeds. We show that fine-tuning can have a significant impact on alignment, depending mainly on the downstream task and the model. Finally, we show that realignment can, in some instances, improve cross-lingual transfer, and we identify conditions in which realignment methods provide significant improvements. Namely, we find that realignment works better on tasks for which alignment is correlated with cross-lingual transfer when generalizing to a distant language and with smaller models, as well as when using a bilingual dictionary rather than FastAlign to extract realignment pairs. For example, for POS-tagging, between English and Arabic, realignment can bring a +15.8 accuracy improvement on distilmBERT, even outperforming XLM-R Large by 1.7. We thus advocate for further research on realignment methods for smaller multilingual models as an alternative to scaling.

1 Introduction

With the more general aim of improving the understanding of Multilingual Large Language Models (MLLM), we study the link between the multilingual alignment of their representations and their ability to perform cross-lingual transfer learning, and investigate conditions for realignment methods to improve cross-lingual transfer.

MLLMs, like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a), are Transformer encoders (Vaswani et al., 2017) which show an effective ability to perform Cross-lingual Transfer Learning (CTL). Despite the absence of any explicit cross-lingual training signal, mBERT and XLM-R can be fine-tuned on a specific task in one language and then provide high accuracy when evaluated on another language on the same task (Pires et al., 2019; Wu and Dredze, 2019). By alleviating the need for training data for a specific task in all languages and for translation data which more than often lacks for non-English languages, CTL with MLLMs could help bridge the gap in NLP between English and other languages.

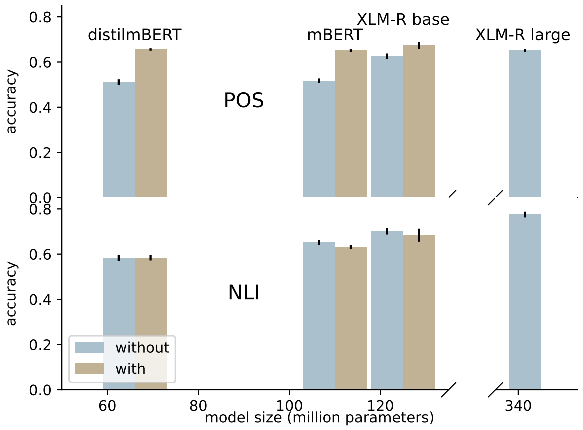


Figure 1: Cross-lingual transfer between English and Arabic with and without realignment, using a bilingual dictionary. For some tasks, realignment can make small models competitive with a large baseline.

But the ability of MLLMs to generalize across languages is highly correlated with the similarity between the training language (often English) and the language to which we hope to transfer knowledge (Pires et al., 2019; Wu and Dredze, 2019). For distant and low-resources languages, CTL with mBERT can give worse results than fine-tuning a Transformer from scratch (Wu and Dredze, 2020a).

Realignment methods (Wu and Dredze, 2020b), sometimes called adjustment or explicit alignment, aim to improve the cross-lingual properties of an MLLM by trying to make similar words from dif-

Findings of the Association for Computational Linguistics: ACL 2023, pages 3020–3042
July 9-14, 2023 ©2023 Association for Computational Linguistics

ferent languages have closer representations. Realignment methods typically require a translation dataset and an alignment tool, like FastAlign (Dyer et al., 2013), to extract contextualized pairs of translated words that will be realigned.

Despite some encouraging results on specific tasks, current realignment methods might not consistently improve cross-lingual zero-shot abilities of mBERT and XLM-R (Wu and Dredze, 2020b). When tested with several seeds on various fine-tuning tasks, improvements brought by realignment are not always significant and do not compare with the gain brought by scaling the model, e.g., from XLM-R Base to XLM-R Large. However, these realignment methods were not tried on smaller models like distilmBERT as we do here.

The mitigated results of realignment methods raise the question of whether cross-lingual transfer is at all linked with multilingual alignment. If improving alignment does not necessarily improve CTL, then the two might not be correlated. Despite the ability of mBERT and XLM-R to perform CTL, there lacks consensus on whether they actually hold aligned representations (Gaschi et al., 2022).

We thus investigate the link between alignment and CTL, with three contributions: (1) We find a high correlation between multilingual alignment and cross-lingual transfer for multilingual Transformers, (2) we show that, depending on the downstream task, fine-tuning on English can harm the alignment to different degrees, potentially harming cross-lingual transfer, and (3) we link our findings to realignment methods and identify conditions under which they seem to bring the most significant improvements to zero-shot transfer learning, particularly on smaller models as shown on Fig. 1.

2 Related Work

Current realignment methods are applied on a pre-trained model before fine-tuning in one language (typically English). Common tasks are Natural Language Inference (NLI), Named Entity Recognition (NER), Part-of-speech tagging (POS-tagging) or Question Answering (QA). The model is then expected to generalize better to other languages for the task than without the realignment. Realignment methods rely on pairs of words extracted from translated sentences using a word alignment tool, usually FastAlign (Dyer et al., 2013), but other tools like AWESOME-align (Dou and Neubig, 2021) could be used. Various realignment objectives are used to bring closer together the contextualized embeddings of words in such pairs: using a linear mapping (Wang et al., 2019), a ℓ_2-based loss with regularization to avoid degenerative solutions (Cao et al., 2020; Zhao et al., 2021), or a contrastive loss (Pan et al., 2021; Wu and Dredze, 2020b).

Existing realignment methods might not significantly improve cross-lingual transfer. Despite improvements on NLI (Cao et al., 2020; Zhao et al., 2021; Pan et al., 2021) or on dependency parsing (Wang et al., 2019), the results might not hold across tasks and languages. A comparative study by Kulshreshtha et al. (2020) showed that methods based on linear mapping are effective only on "moderately close languages", whereas ℓ_2-based loss improves results for "extremely distant languages". This latter ℓ_2-loss was shown to work well on a NLI task, but not for all languages on a NER task, and to be even detrimental for QA tasks (Efimov et al., 2022). Finally, Wu and Dredze (2020b) compared linear mapping realignment, ℓ_2-based realignment and contrastive learning on several tasks, languages and models, performing several runs. They found that existing methods do not bring consistent improvements over no realignment.

Expecting realignment methods to succeed implies a direct link between the multilingual alignment of the representations produced by a model and its ability to perform CTL. However, there isn't any strong consensus on whether multilingual Transformers have well-aligned representations (Gaschi et al., 2022), let alone on whether better-aligned representations lead to better CTL.

Assessing the multilingual alignment of contextualized representations can take many forms. Pairs of words are extracted from translated sentences, usually with FastAlign or a bilingual dictionary (Gaschi et al., 2022). Then, after building contextualized representations of the words of each pair, the distribution of their similarity can be compared with that of random pairs of words (Cao et al., 2020). But this method can lead to incorrect conclusions (Efimov et al., 2022). A high overlap in the distribution of similarities between related and random pairs means that sometimes random pairs can have higher similarities than related pairs. But since those pairs do not necessarily involve the same words, a high overlap does not mean that any word is closer to an unrelated one than to a related one. An alternative is to compare a related pair to its neighbors (Efimov et al., 2022), which shows that realignment methods indeed improve multilingual alignment and that fine-tuning can harm this alignment. Another similar approach consists in designing a nearest-neighbor search criterion. This was done for sentence-level representations (Pires et al., 2019) and for word-level alignment (Conneau et al., 2020b; Gaschi et al., 2022), showing that MLLMs like mBERT have a multilingual alignment that is competitive with static embeddings (Bojanowski et al., 2017) explicitly aligned with a supervised alignment method (Joulin et al., 2018).

3 Method

To study the link between multilingual alignment and cross-lingual transfer (CTL), we need a way to evaluate alignment and CTL. We use a relative difference to evaluate CTL, we discuss different methods for evaluating alignment, and describe the realignment method used in our experiments.

3.1 Evaluating cross-lingual transfer

A model has high CTL abilities when, after fine-tuning for one language, it can obtain a high evaluation score on other languages. To evaluate it for a given task, we compute the relative difference between the evaluation metric m_{en} on the English development set and the evaluation metric m_{tgt} on the target language:

$$\text{cross-lingual transfer} = \frac{m_{\text{tgt}} - m_{\text{en}}}{m_{\text{en}}} \qquad (1)$$

The monolingual metric is a score between 0 and 1, like accuracy or f1-score, where higher is better. Then our metric gives scores between -1 and $+\infty$. A negative score is obtained if and only if $m_{\text{tgt}} < m_{\text{en}}$, which should always be the case in practice. Values closer to 0 then indicate better CTL for a specific task and language.

It must be noted that for datasets where the target language test set is a translation of the English one, the normalization in Equation 1 allows the metric to boild down roughly to minus the proportion of correct answers in English that were misclassified when translated, assuming there isn't not to many misclassified English examples that were correctly classified in the target language, which should be the case since there are not that much misclassified English examples in general.

3.2 Evaluating alignment

To evaluate multilingual alignment, we use the same method for extracting pairs of translated words with their context as Gaschi et al. (2022). Provided a source of related pairs of words from both languages, a fixed number of pairs of words are randomly selected and a nearest-neighbor search with cosine similarity is performed. The top-1 accuracy of the nearest-neighbor search is the score of the alignment evaluation.

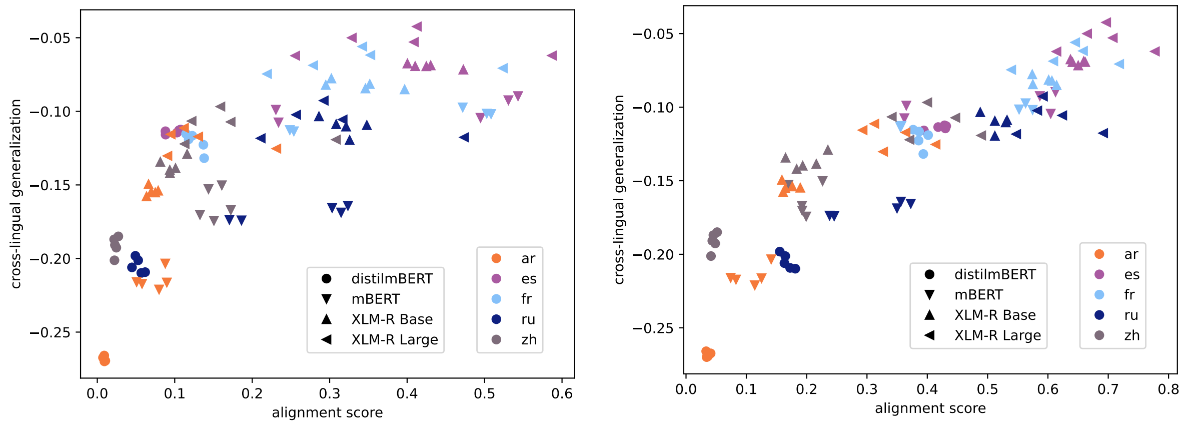
To extract contextualized pairs of translated words from a translation dataset, FastAlign is the most widely used word aligner in realignment methods (Wu and Dredze, 2020b; Cao et al., 2020; Zhao et al., 2021; Wang et al., 2019), but it is prone to errors and thus generates noisy training realignment data (Pan et al., 2021; Gaschi et al., 2022). Following Gaschi et al. (2022), we use a bilingual dictionary to extract matching pairs of words in translated sentences, discarding any ambiguity to obtain the most accurate pairs possible.

It is worth noting that the accuracy of a nearest-neighbor search is not symmetric. We use the convention that an A-B alignment means that we look for the translation of each word of language A among its nearest neighbors. Two types of alignment can be evaluated: strong and weak alignment (Roy et al., 2020). Weak alignment is the expected way to compute alignment: when evaluating A-B weak alignment, we search a translation for a given word of A only among nearest-neighbors belonging to B. But with such an evaluation, there can be situations with highly measured alignment where representations from both languages are far apart with respect to intra-language similarity. Strong alignment remedies to this by including language A in the search space. With A-B strong alignment, we search a translation for a given word of A among its nearest-neighbors belonging to *both* language B *and* A. For a given pair of related words to be considered close enough, the word from language B must be closer to its translation in A than any other word from B *and* A. We show in our experiments that strong alignment is more correlated with CTL than weak alignment.[1]

3.3 Realignment loss

A realignment task consists in making the representations of related pairs closer to each other. The method used to extract related pairs for alignment evaluation can be used for computing the realignment loss. Following Wu and Dredze (2020b), we

[1] Although strong alignment can be affected by synonyms, restricting the search space to the words from the sampled extracted pairs reduces the risk of founding a synonym.



(a) NLI, last layer

(b) NLI, penultimate

Figure 2: Plot of CTL abilities against the English-target strong alignment measured for the last and penultimate layer after fine-tuning on NLI.

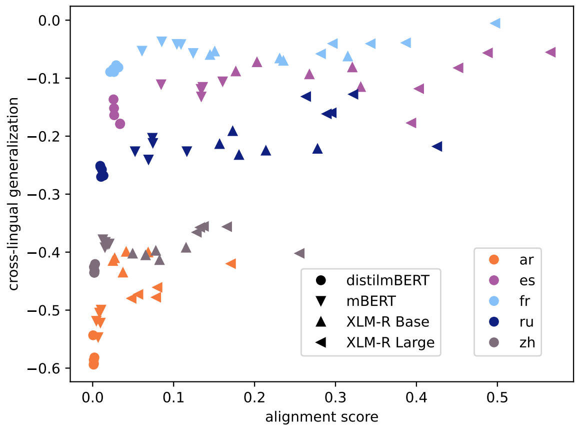


Figure 3: CLT abilities against English-target strong alignment for the last layer after fine-tuning on NER.

minimize a contrastive loss using the framework of Chen et al. (2020), encouraging strong alignment for pairs within a batch. A batch is composed of a set of representations $\mathcal{H}$ of all words in a few pairs of translated sentences and a set $\mathcal{P} \subseteq \mathcal{H} \times \mathcal{H}$ containing the pairs of translated words extracted with a bilingual dictionary (or a word aligner). The realignment loss can then be written as:

$$\mathcal{L}_{\text{realign}} = -\frac{1}{2|\mathcal{P}|}\sum_{(s,t)\in\mathcal{P}}\left[\log\frac{\exp\left(\frac{\text{sim}(s,t)}{T}\right)}{\sum_{\substack{h\in\mathcal{H}\\ h\neq s}}\exp\left(\frac{\text{sim}(s,h)}{T}\right)} + \log\frac{\exp\left(\frac{\text{sim}(s,t)}{T}\right)}{\sum_{\substack{h\in\mathcal{H}\\ h\neq t}}\exp\left(\frac{\text{sim}(h,t)}{T}\right)}\right] \tag{2}$$

For each pair (s,t), the cosine similarity (sim) is compared to negative pairs (all other pairs in $\mathcal{H}$) using a softmax with a temperature hyper-parameter ($T = 0.1$), following Wu and Dredze (2020b), bringing closer together translated pairs of words with respect to other pairs in the batch.

	POS	NER	XNLI
en-train	12,543	20,000	392,703
en-dev	2,002	10,000	2,490
en-test	2,077	10,000	5,010
ar-test	680	10,000	5,010
es-test	426	10,000	5,010
fr-test	416	10,000	5,010
ru-test	601	10,000	5,010
zh-test	500	10,000	5,010

Table 1: Number of examples

3.4 Experimental details

We evaluate cross-lingual transfer with three multilingual tasks, the sizes of which are reported in Table 1:

- Part-of-speech tagging (POS-tagging) with the Universal Dependencies dataset (Zeman et al., 2020). Similarly to Wu and Dredze (2020b), we use the following treebanks: Arabic-PADT, English-EWT, Spanish-GSD, French-GSD, Russian-GSD, and Chinese-GSD.

- Named Entity Recognition (NER) with the WikiANN dataset (Pan et al., 2017).

- Natural Language Inference (NLI) with the XNLI dataset (Conneau et al., 2018).

It must be noted that XNLI is the only dataset with translated test sets, and thus the only one for

task	layer	weak		strong	
		before	after	before	after
POS	last	0.58	0.84	0.82	0.87
	penult	0.78	0.84	0.87	0.86
NER	last	0.69	0.72	0.86	0.70
	penult	0.82	0.71	0.87	0.82
NLI	last	0.51	0.75	0.86	0.82
	penult	0.74	0.95	0.84	0.92

Table 2: Spearman's rank correlation of CTL with the English-target alignment produced by the last and penultimate layer before and after fine-tuning. Evaluation is done across 5 languages, 5 seeds and 4 models ($N = 100$). All cells have p-value < 0.05.

which the cross-lingual transfer metric is strictly comparable across languages. In our experiments, high correlation will nonetheless be observed between CTL and alignment for the two other tasks, suggesting that the CTL metrics is not so much affected by difference in size and domain between the test sets.

Further details about implementation can be found in Appendix B And in the source code[2].

4 Correlation between alignment and CTL

We measure the correlation between multilingual alignment and cross-lingual transfer (CTL) across models, languages and seeds. We also compare the correlation between alignment before fine-tuning and after fine-tuning with CTL and with different alignment measures.

Spearman's rank correlation is measured between alignment before or after fine-tuning and CTL. The English-target alignment is computed for each target language with the method described in Section 3.2 and is compared with the transfer ability from English to that same target language with the metric described in Section 3.1.

Table 2 shows correlations between CTL and different types of alignment. It is computed separately for each different task (POS, NER, NLI), for the alignment at the last and second to last layer (last and penult), before and after fine-tuning on the given task, and with weak and strong alignment. Comparing other layers for models of different sizes is less relevant, since the correlation is computed across models with various number of layers. And a model-by-model analysis of the correlation with the alignment in various layers did not reveal contradictory results (cf. Appendix E). Each correlation value is obtained from 100 samples with four different models (distilmBERT, mBERT, XLM-R Base and Large), five target languages (Arabic, Spanish, French, Russian and Chinese) and five seeds for initialization of the classification head and shuffling of the fine-tuning data.

Results show that strong alignment is better correlated to cross-lingual transfer than weak alignment. With the exception of two tasks after fine-tuning (NER and NLI), strong alignment has a marginally higher correlation with CTL. This is particularly noticeable when looking at alignment before fine-tuning on the last layer, going from a correlation between 0.51 and 0.69 for weak alignment to one ranging from 0.82 to 0.86 for strong alignment.

Tab. 2 also shows that for NLI, the alignment on the penultimate layer seems better correlated to cross-lingual transfer than with the last layer. A relatively important gap in correlation is measured between the last and the second-before-last layer for all cases except for strong alignment before fine-tuning. The fact that alignment on the penultimate layer would correlate better than the last for NLI can be explained by the sentence-level nature of the task. For sentence classification tasks, the classification head reads only the representation of the first token of the last layer, which is computed from the representations of all the tokens at the previous layer, leading to a pooling of the penultimate layer.

Despite the different values observed, there seems to be no significant difference between correlation for alignment measured before and after fine-tuning, and a careful analysis of confidence interval obtained with bootstrapping (Efron and Tibshirani, 1994) can confirm this (cf. Appendix C for detailed results).

Fig. 2 shows the relation between CTL and English-target strong alignment measured after fine-tuning measured in four situations to further illustrate the link between alignment and transfer.

Fig. 2b shows one of the cases with higher correlation (0.92). The correlation seems to hold well across models (forms) and languages (colors). However, for a given model and language, the random seed for fine-tuning seems to be detrimental to the correlation, although at a small scale. Hence, alignment might not be the only factor to affect cross-lingual generalization as the model initializa-

[2] https://github.com/posos-tech/multilingual-alignment-and-transfer

tion or the data shuffling seems to play a smaller role.

Fig. 3 shows a case with one of the lowest correlations between strong alignment and CTL (0.70). It seems that models and initialization seeds have a higher impact on alignment than on CTL. For example, in the case of English-French alignment (green), CTL is between 0.0 and -0.1, whatever the model and seed, not overlapping with other target-English language pairs, but alignment varies between approximately 0.05 and 0.5, overlapping with all other language pairs. Interestingly, the penultimate layer has a higher accuracy (0.82), suggesting that for NER the last layer is not necessarily the one for which alignment correlates the most with CTL.

For two of the three tested tasks (NER and POS-tagging), it must be noted that the CTL metric is not strictly comparable across languages since the test sets for each language are of different domains and sizes (cf. Section 3.4). However, for the third task (NLI), each test set is a translation of the English one, and thus the CTL metric is strictly comparable in that case. This might explain why correlations are higher for the NLI task than the others. Nevertheless, the observed correlation for the two other tasks is still significantly high, which suggests that the general tendency might not be affected by the differences in domains and sizes in the test sets.

5 The impact of fine-tuning on alignment

To study the link between alignment and cross-lingual transfer (CTL), we also look at the impact of fine-tuning over alignment. We've already shown that strong alignment is highly correlated with CTL. However, we weren't able to conclude whether alignment measured before or after fine-tuning was better correlated to CTL abilities. To understand the difference between both measures, we study in this section the impact of fine-tuning on the alignment of MLLMs representations. We use the same fine-tuning runs as in the previous section (4).

Tab. 3 shows the relative variation in alignment before and after fine-tuning for all tasks and models tested and for three languages for clarity (complete results in Appendix D). The relative difference is built in the same way as the cross-lingual transfer evaluation (Eq. 1). Negative values indicate a drop in alignment. Alignment is measured at the last layer. Fig. 4 and 5 show a breakdown by layer for a few cases.

task	model	en-ar	en-es	en-ru
POS	distilmBERT	-0.74	-0.86	-0.87
	mBERT	-0.90	-0.86	-0.95
	XLM-R Base	-0.43	-0.46	-0.70
	XLM-R Large	-0.30	0.23	-0.44
NER	distilmBERT	0.00	-0.61	-0.33
	mBERT	-0.28	-0.36	-0.27
	XLM-R Base	5.88	0.22	1.32
	XLM-R Large	16.34	2.22	3.10
NLI	distilmBERT	5.49	0.30	2.28
	mBERT	5.65	0.99	1.45
	XLM-R Base	11.17	1.01	2.67
	XLM-R Large	25.36	1.78	2.99

Table 3: Relative variation of strong alignment at the last layer before and after fine-tuning for different fine-tuning tasks (nearest-neighbor search accuracy after fine-tuning minus accuracy before).

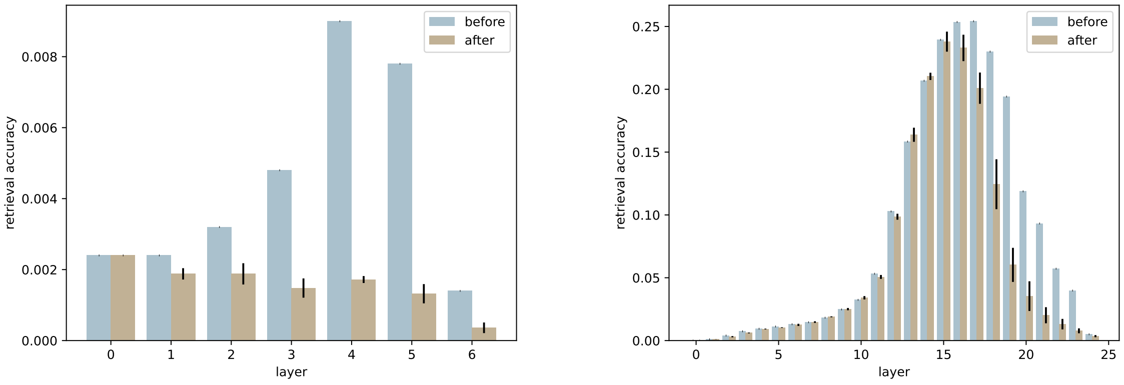
For certain combinations of models and tasks, fine-tuning is detrimental to multilingual alignment. distilmBERT and mBERT mainly show a decrease in alignment for POS-tagging and NER, and smaller improvements than other models on NLI. However, POS-tagging is the only of the three tasks which shows dramatic drops where alignment can be reduced by as much as 96%.

The drop in alignment can be explained by catastrophic forgetting. If the model is only trained on a monolingual task, it might not retain information about other languages or about the link between English and other languages.

What is more surprising is the increase in alignment obtained in other cases. XLM-R Base and Large, which are larger models than mBERT and distilmBERT, have a relative increase that can go as high as 25.36 on the NLI task for distant languages. And although these increases are from a small alignment measure, we still observe a large increase for middle layers where the initial alignment is already quite high (cf. Fig. 5).

The alignment of larger models being less harmed by fine-tuning is coherent with the fact that those same larger models have been shown to have better CTL abilities. Fig. 4 shows that more layers seem to mitigate the potentially negative impact of fine-tuning on alignment, as it affects mainly the layers closest to the last one and as the initial alignment measure is globally higher for XLM-R than distilmBERT (before fine-tuning: ≈0.25 against ≈0.008).

Giving a definitive answer as to why different tasks have different impacts on alignment might need further research. But one could already argue



(a) distilmBERT POS

(b) XLM-R Large POS

Figure 4: Evolution across layers of English-Arabic alignment before and after fine-tuning of distilmBERT and XLM-R Large on POS-tagging, starting at 0 for the embedding layer.

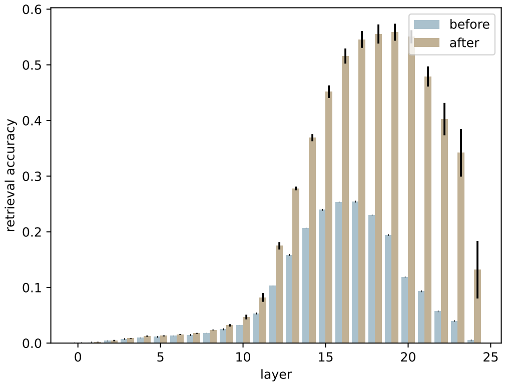


Figure 5: Alignment across layers of English-Arabic before and after fine-tuning of XLM-R Large on NLI.

that each task corresponds to different levels of abstraction in NLP. Tasks with a low level of abstraction like POS-tagging might rely on the word form itself and thus on more language-specific components of the representations, which when enhanced, decreases alignment. On the other hand, NLI has a higher level of abstraction, requiring the meaning rather than the word form, which might be encoded in deeper layers (Tenney et al., 2019) which are more aligned.

Fine-tuning MLLMs on a downstream task has an impact on the multilingual alignment of the representations produced by the model. For "smaller" language models, it is systematically detrimental, as well as for certain tasks like POS-tagging. This might explain why some realignment methods might not work for all models nor all tasks (Wu and Dredze, 2020b).

6 Impact of realignment on cross-lingual transfer

We have already shown that the correlation between multilingual alignment and cross-lingual transfer (CTL) is high (Section 4). But we do not know whether they are more directly linked. In this section, we try to identify the conditions under which improving alignment in multilingual models leads to improvement in CTL.

Sequential realignment is the usual way to perform realignment: realignment steps are performed on the pre-trained model before fine-tuning. We propose to compare it with joint alignment, where we optimize simultaneously for the realignment and the downstream task (more details in Appendix A), to try and identify whether alignment before or after fine-tuning is more strongly related to CTL.

In the same settings as the previous experiments (tasks, models and languages, and number of seeds), we fine-tune models in English with different realignment methods and evaluate CTL on different languages. Following a similar setting as (Wu and Dredze, 2020b), realignment data from the five pairs of languages (English-target) is interleaved to form a single multilingual realignment dataset. Models are fine-tuned on POS-tagging or NER for five epochs and 2 epochs for NLI because its training data is larger. We use the opus100 translation dataset (Zhang et al., 2020) from which we extract pairs of words using bilingual dictionaries. We also tested with the multiUN translation data (Ziemski et al., 2016), which conditioned our choice of languages, and with other ways to extract alignment pairs: FastAlign (Dyer et al., 2013)

	POS	NER	NLI
distilmBERT	*73.1*	*57.4*	*64.2*
+ before	**78.0**	58.9	64.3
+ joint	77.9	**59.6**	**65.0**
mBERT	*74.3*	*62.2*	*69.7*
+ before	78.2	62.7	68.7
+ joint	**78.3**	**64.8**	**70.0**
XLM-R base	*78.8*	*60.4*	*74.0*
+ before	**79.9**	**63.9**	72.7
+ joint	79.4	63.0	**74.6**
XLM-R large	*79.6*	*65.0*	*80.0*

Table 4: Condensed results of the controlled experiment comparing joint and sequential realignment using a bilingual dictionary. Light gray indicates a difference with baseline lower than its standard deviation. Dark gray indicates lower than baseline minus standard deviation.

POS	ar	es	fr	ru	zh
distilmBERT	*51.0*	*84.1*	*85.3*	*81.2*	*64.1*
+ before	65.5	**86.5**	**85.8**	**84.7**	**67.4**
+ joint	**66.8**	**85.8**	86.5	84.1	66.4
XLM-R base	*62.5*	*86.6*	*86.9*	***86.9***	*70.9*
+ before	**67.3**	**86.9**	**87.3**	86.8	**71.2**
+ joint	66.6	86.6	87.2	86.0	70.6
XLM-R large	***65.1***	***87.0***	***87.5***	***87.0***	***71.5***

Table 5: Breakdown of realignment results for some languages and distilmBERT and XLM-R.

and AWESOME-align (Dou and Neubig, 2021). Changing the translation dataset does not fundamentally change the results, and using probabilistic alignment tools made realignment methods less effective. The results presented in this section were handpicked for the sake of clarity, but the reader can refer to Appendix F.

Condensed results are reported on Tab. 4, averaged on the five languages. A breakdown by languages for the POS-tagging task and two models is shown on Tab. 5. It shows that realignment methods improve performance only on certain tasks, models and language pairs.

Realignment methods, either sequential or joint, provide significant improvement for all models for the POS-tagging task, but less significant ones for NER, and no significant improvement for NLI. The positive impact of realignment on cross-lingual transfer seems to be mirrored by the negative impact of fine-tuning over alignment. Indeed, POS-tagging is also the task for which fine-tuning is the most detrimental to multilingual alignment, as shown in the previous section.

The same parallel can be drawn for models. distilmBERT is the model that benefits the most from realignment. It is also the one whose alignment suffers the most from fine-tuning. Smaller multilingual models seem to benefit more from realignment, as well as they see their multilingual alignment reduced after fine-tuning. In the same way that fine-tuning mainly affects the deeper layers, it is possible that realignment might affect only those deeper layers. This would mean that most layers would have their alignment significantly improved for small models like distilmBERT (6 layers), while larger models might be only superficially realigned.

Finally, besides tasks and models, it can also be observed that the impact of realignment varies across language pairs (Tab. 5). Although we did not test on many language pairs, results are coherent with the idea that realignment methods tend to work better on distant pairs of languages (Kulshreshtha et al., 2020).

On a side note, our controlled experiment does not allow us to conclude whether it is more important to improve alignment before fine-tuning or after. It seems that alignment measured before and after fine-tuning are equally important to cross-lingual transfer.

Realignment methods unsurprisingly provide better results when the alignment is lower, be it before or after fine-tuning. Distant languages and small models have lower alignment, and POS-tagging is a task where alignment decreases after fine-tuning. Realignment helps only up to a certain point where representations are already well aligned, and CTL gives already good results. For distilmBERT on POS-tagging for transfer from English to Arabic, it provides a +15.8 improvement over baseline, even outperforming XLM-R Large by 1.7 points. In such conditions, realignment is an interesting alternative to scaling for multilingual models.

If realignment succeeds in some favorable conditions, then how can we explain that realignment methods were shown to not be significantly improving CTL on several tasks, including POS-tagging (Wu and Dredze, 2020b)? Firstly, to the best of our knowledge, realignment was never tried on distilmBERT or other models of equivalent size. Secondly, Tab. 6 shows that it might be partly due to an element of the realignment methods that was overlooked: the source of related pairs of words.

The way pairs are extracted seems to be crucial to the success of realignment methods. Tab. 6 shows the effect of different types of pairs extraction in realignment methods. Realignment methods using pairs extracted with FastAlign or

	POS	NER
XLM-R base	*78.8*	*60.4*
+ before fastalign	78.6	61.4
+ before awesome	78.6	62.0
+ before dico	**79.9**	**63.9**
+ joint fastalign	78.0	62.1
+ joint awesome	77.8	62.3
+ joint dico	79.4	63.0

Table 6: Average CTL abilities for XLM-R with different type of realignment.

AWESOME-align do not provide significant improvements over the baseline, whereas using a bilingual dictionary does. Using a bilingual dictionary might be more accurate for extracting translated pairs (Gaschi et al., 2022). Another explanation could be that the type of words contained in a dictionary might help since it might contain more lexical words holding meaning and fewer grammatical words.

7 Conclusion

We have shown that multilingual alignment, measured using a nearest-neighbor search among translated pairs of contextualized words, is highly correlated with the cross-lingual transfer abilities of multilingual models (or at least multilingual Transformers). Strong alignment was also revealed to be better correlated to cross-lingual transfer than weak alignment.

Then we investigated the impact of fine-tuning (necessary for cross-lingual transfer) on alignment as well as the impact of realignment methods on cross-lingual transfer. Fine-tuning was revealed to have a very different impact on alignment depending on the downstream task and the model. Where lower-level tasks seemed to have the most impact and smaller models seemed to be the most affected. Conversely, realignment methods were shown to work better on those same tasks and models. Ultimately, realignment works unsurprisingly better when the baseline alignment (before or after fine-tuning) is lower.

We also showed that using a bilingual dictionary for extracting pairs for realignment methods improves over the commonly used FastAlign and over a more precise neural aligner (AWESOME-align).

It's worth noting that realignment works particularly well for a small model like distilmBERT (66M parameters), allowing it in some cases to obtain competitive results with XLM-R Large (354M parameters). This advocates for further research on realignment for small Transformers to build more compute-efficient multilingual models.

Finally, further research is needed to investigate additional questions, like whether cross-lingual transfer is more directly linked to alignment before or after fine-tuning, or to alignment at certain layers for certain tasks. To answer these questions, more large-scale experiments could be performed on more tasks and especially on more languages to obtain correlation values with smaller confidence intervals.

8 Limitations

We worked with only five language pairs, all involving English and another language: Arabic, Spanish, French, Russian and Chinese. This is due to using the multiUN dataset (Ziemski et al., 2016) for evaluating alignment and performing realignment. We also used the opus100 dataset (Zhang et al., 2020), which contains more pairs and is the dataset that eventually figured in our paper, but we stuck to the same language pairs for a fair comparison with multiUN in Appendix F. This narrow choice of language limits our ability to understand why realignment methods work well for some languages and not others. And we believe that making a similar analysis with many language pairs, not necessarily involving English, would be a good lead for further research investigating the link between the success of the realignment method and how two languages relate to each other.

We chose a strong alignment objective with contrastive learning for our realignment task. Several other objectives could have been tried, like learning an orthogonal mapping between representations (Wang et al., 2019) or simply using a ℓ_2-loss to collapse representations together (Cao et al., 2020), but both methods require an extra regularization step (Wu and Dredze, 2020b) since they do not leverage any negative samples. For the sake of simplicity, we focused on a contrastive loss, as trying different methods would have led to an explosion in the number of runs for the controlled experiment. This also explains why we used the same hyper-parameters and pre-processing steps of Wu and Dredze (2020b). A more thorough search for the optimal parameters, and realignment loss, might lead to better results.

9 Acknowledgements

We would like to thank the anonymous reviewers for their comments, as well as Shijie Wu, who kindly explained some details in the implementation of his paper Wu and Dredze (2020b). We are also grateful for discussion and proof-reading brought by our colleagues at Posos: François Plesse, Xavier Fontaine and Baptiste Charnier.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations[3].

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Anthony J. Bishara and James B. Hittner. 2017. Confidence intervals for correlations when data are not normal. *Behavior Research Methods*, 49(1):294–309.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

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

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. 2018. Adapting auxiliary losses using gradient similarity.

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.

Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. 2022. The impact of cross-lingual adjustment of contextual word representations on zero-shot transfer.

Bradley Efron and Robert Tibshirani. 1994. An introduction to the bootstrap.

Félix Gaschi, François Plesse, Parisa Rastin, and Yannick Toussaint. 2022. Multilingual transformer encoders: a word-level task-agnostic evaluation.

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(7825):357–362.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a

[3] see https://www.grid5000.fr

retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

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

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lukas Liebel and Marco Körner. 2018. Auxiliary tasks in multi-task learning. *CoRR*, abs/1805.06334.

Fenglin Liu, Meng Gao, Yuanxin Liu, and Kai Lei. 2019. Self-adaptive scaling for learnable residual structure. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 862–870, Hong Kong, China. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. 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, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

John Ruscio. 2008. Constructing confidence intervals for spearman's rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods. *Journal of Modern Applied Statistical Methods*, 7:7.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020a. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020b. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

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 Hồng, 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 Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Sheng Zhang, Kevin Duh, and Benjamin Van Durme. 2018. Fine-grained entity typing through increased discourse context and adaptive classification thresholds. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 173–179, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Joint realignment

Existing realignment methods proceed in a sequential manner. The pre-trained model is first optimized for the realignment loss, before any fine-tuning. This assumes that the alignment before fine-tuning is positively linked to the cross-lingual transfer abilities of the model and that improving alignment before fine-tuning will improve transfer. However, fine-tuning itself might have an impact on alignment (Efimov et al., 2022).

To compare the importance of alignment before and after fine-tuning for CTL, we introduce a new realignment method where realignment and fine-tuning are performed jointly. We optimize simultaneously for a realignment loss and the fine-tuning loss. In practice, for each optimization step, we compute the loss $\mathcal{L}_{\text{task}}$ for a batch of the fine-tuning task and the loss $\mathcal{L}_{\text{realign}}$ for a batch of the alignment data. The total loss for each backward pass is then written as:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{realign}} \tag{3}$$

This joint realignment can be framed as multi-task learning. The fine-tuning task would be the main task and the realignment task an auxiliary one. There are more elaborate methods for training a model with an auxiliary task (Liebel and Körner, 2018; Du et al., 2018; Zhang et al., 2018; Liu et al., 2019) but our aim is to propose the simplest method possible to compare joint and sequential alignment in a controlled setting.

B Experimental details

B.1 Scientific artifacts used

We relied on the following scientific Python packages for our experiments: the HuggingFace's libraries `transformers` (Wolf et al., 2020), `datasets` (Lhoest et al., 2021) and `evaluate`[4], `PyTorch` (Paszke et al., 2019), `NLTK` (Bird et al., 2009) and its implementation of the Stanford Chinese Segmenter (Tseng et al., 2005), `seqeval` (Nakayama, 2018) for evaluating NER, `NumPy` (Harris et al., 2020), and AWESOME-align (Dou and Neubig, 2021), FastAlign (Dyer et al., 2013), and MUSE dictionaries (Lample et al., 2018) for extracting alignment pairs.

We used the following datasets: two translation datasets for building evaluating alignment and realigning, multiUN (Ziemski et al., 2016) and opus100 (Zhang et al., 2020), XNLI (Conneau et al., 2018), the Universal Dependencies dataset (Zeman et al., 2020) for POS-tagging, and the WikiANN dataset for NER (Pan et al., 2017).

model	# parameters
distilmBERT	66M
mBERT	110M
XLM-R Base	125M
XLM-R Large	345M

Table 7: Number of parameters

Finally, we worked with four different models: distilmBERT, which was released with distilBERT (Sanh et al., 2019), mBERT, which was released with BERT (Devlin et al., 2019) and XLM-R Base and Large (Conneau et al., 2020a).

B.2 Multilingual alignment data

From a translation dataset, pairs were extracted either using a bilingual dictionary, following (Gaschi et al., 2022), with FastAlign (Dyer et al., 2013) or AWESOME-align (Dou and Neubig, 2021). For FastAlign, we produced alignments in both direction and symmetrize with the `grow-diag-final-and` heuristic provided with FastAlign, following the setting of Wu and Dredze (2020b). For all methods of extraction, we kept only one-to-one alignment and discard trivial cases where both words are identical, again following Wu and Dredze (2020b).

B.3 Experimental setup

We performed two experiments:

1. Fine-tuning all models on all tasks for 5 epochs and measuring alignment before and after fine-tuning. This experiment provided the results for Section 4 and 5.
2. Performing different realignment methods before fine-tuning for 5 epochs (or 2 for XNLI), providing results for section 6.

For both experiments, we reused the experimental setup from Wu and Dredze (2020b). Fine-tuning on a downstream task is done with Adam, with a learning rate of 2e-5 with a linear decay and warmup for 10% of the steps. Fine-tuning is performed on 5 epochs, and 32 batch size, except for XNLI in the second experiment, where we trained for 2 epochs, which still leads to more fine-tuning steps than any of the two other tasks (cf. Table 1).

[4] `https://huggingface.co/docs/evaluate/index`

For the realignment methods, still following Wu and Dredze (2020b), we train in a multilingual fashion, where each batch contains examples from all target languages. However, we use the same learning rate and schedule as for fine-tuning for a fair comparison between joint and sequential realignment, since the same optimizer is used for fine-tuning and realignment when performing joint realignment. We use a maximum length of 96 like Wu and Dredze (2020b) but a batch size of 16 instead of 128 because of limited computing resources.

B.4 Discussion on the number of realignment samples

It is worth noting that our method uses fewer realignment samples. Since we alternate batches of 16 realignment samples and batches of 32 fine-tuning samples for joint realignment, this fixes the number of realignment samples we will use for a specific downstream task, for a fair comparison. This gives 31,358 sentence pairs for POS-tagging, 50,000 for NER, and 392,703 for NLI. For comparison, Wu and Dredze (2020b) used 100k steps of batches of size 128. The number of realignment samples used could have been a factor explaining why realignment works well for POS-tagging and less for NER and NLI, and why Wu and Dredze (2020b) do not find that realignment methods improve results significantly on any task. It could be argued that training on too many realignment samples might hurt performances. However, when testing on the POS-tagging task, we found that the number of realignment samples did not have significant impact on performances.

B.5 Computational budget

The first experiment was performed on Nvidia A40 GPUs for an equivalent of 3 days for a single GPU (including all models, tasks and seeds). For the second experiment, training (fine-tuning and/or realignment) was performed on various smaller GPUs (RTX 2080 Ti, GTX 1080 Ti, Tesla T4) for distilmBERT, mBERT and XLM-R Base, and on a Nvidia A40 for XLM-R Large. The experiment took more than 10 GPU-days on the smaller GPUs, combining all models, realignment methods (including baseline), random seeds, translation datasets and pairs extraction methods. For XLM-R Large, for which we only trained the baseline, it still required 30 GPU-hours on Nvidia A40.

C Confidence intervals for correlation

In Section 4 we compared correlation for different tasks, before and after fine-tuning, for English-target and target-English alignment and for the last and penultimate layer. These correlations where computed across several models, languages and seeds. From this correlation statistics, we have drawn three conclusions:

1. Strong alignment is better correlated with cross-lingual transfer than weak alignment.
2. The NLI task, because of its sentence-level nature, have a cross-lingual transfer that correlates better with the penultimate layer than the last one.
3. The results do not significantly attribute higher correlation of cross-lingual transfer with alignment before or after fine-tuning neither with English-target compared to target-English alignment.

We verify here that these conclusions hold when looking at the confidence intervals (Tab. 8 and Tab. 9). Confidence intervals are obtained using the Bias-Corrected and Accelerated (BCA) bootstrap method, where several subsets (2000) subsets of our 100 points for each measure of the correlation coefficient are sampled to obtain an empirical distribution of the correlation from which the confidence interval can be deduced (Efron and Tibshirani, 1994). Since we are dealing with ordinal data (the rank in Spearman's rank correlation), bootstrap confidence intervals are expected to have better properties than methods based on assumptions about the distribution (Ruscio, 2008; Bishara and Hittner, 2017)

Is strong alignment significantly better correlated with cross-lingual transfer than weak alignment? comparing both tables cell-by-cell reveals that confidence intervals for the last layer before fine-tuning hardly never overlap, and when they do it's with a small overlap. So in the case of alignment of the last layer before fine-tuning, strong alignment is significantly better correlated with cross-lingual transfer than weak alignment. For other situations, confidence interval overlap. But the fact that strong alignment has almost systematically a higher correlation makes our correlation still relevant.

Does the penultimate layer correlate better than the last one for NLI? For this task, we observe

task	layer	en-X		X-en	
		before	after	before	after
POS	last	0.58 (0.43 - 0.70)	0.84 (0.77 - 0.89)	0.63 (0.48 - 0.74)	0.83 (0.74 - 0.89)
	penult.	0.78 (0.68 - 0.85)	0.84 (0.76 - 0.89)	0.80 (0.71 - 0.87)	0.85 (0.79 - 0.90)
NER	last	0.69 (0.55 - 0.79)	0.72 (0.59 - 0.81)	0.75 (0.64 - 0.83)	0.84 (0.73 - 0.89)
	penult.	0.82 (0.73 - 0.88)	0.71 (0.58 - 0.81)	0.88 (0.83 - 0.92)	0.72 (0.58 - 0.82)
NLI	last	0.51 (0.32 - 0.67)	0.75 (0.61 - 0.85)	0.54 (0.36 - 0.68)	0.73 (0.59 - 0.83)
	penult.	0.74 (0.59 - 0.84)	0.95 (0.90 - 0.97)	0.79 (0.66 - 0.87)	0.94 (0.90 - 0.97)

Table 8: 95% confidence interval for Spearman rank correlation between weak alignment and CTL, obtained with BCA bootstraping with 2000 resamples.

task	layer	en-X		X-en	
		before	after	before	after
POS	last	0.80 (0.73 - 0.86)	0.85 (0.81 - 0.88)	0.83 (0.77 - 0.87)	0.87 (0.83 - 0.91)
	penult.	0.86 (0.79 - 0.89)	0.85 (0.79 - 0.89)	0.87 (0.82 - 0.91)	0.86 (0.82 - 0.90)
NER	last	0.85 (0.78 - 0.90)	0.66 (0.53 - 0.77)	0.86 (0.82 - 0.90)	0.74 (0.65 - 0.82)
	penult.	0.87 (0.83 - 0.91)	0.75 (0.65 - 0.84)	0.88 (0.84 - 0.92)	0.76 (0.66 - 0.84)
NLI	last	0.74 (0.63 - 0.82)	0.81 (0.72 - 0.87)	0.89 (0.83 - 0.93)	0.84 (0.77 - 0.90)
	penult.	0.84 (0.79 - 0.88)	0.92 (0.87 - 0.95)	0.90 (0.86 - 0.93)	0.94 (0.90 - 0.96)

Table 9: 95% confidence interval for Spearman rank correlation between strong alignment and cross-lingual transfer, obtained with BCA bootstraping with 2000 resamples.

that the confidence intervals of the penultimate and last layer do not overlap when the alignment is measured after fine-tuning. Otherwise, before fine-tuning, we can still observe that the measured correlation for the penultimate layer is systematically above the confidence interval for the last layer, except for target-English strong alignment.

We can see that confidence intervals overlap too much when comparing before and after fine-tuning, except in two cases. When looking at POS-tagging for the last layer, weak alignment after fine-tuning gives a significantly better correlation than before, but this does not translate to strong alignment which correlates better with cross-lingual transfer overall. The same observation can be made about NLI for the penultimate layer. On the other hand, for the NER task, strong alignment after-fine tuning gives a significantly worse correlation than before. It is thus difficult to conclude on whether alignment before or after fine-tuning is better correlated to cross-lingual transfer.

Finally, comparing target-English and English-target alignment does not give significant results. If all other parameters are kept identical, every situation leads to an overlap between confidence intervals except for the last layer before fine-tuning for NLI, which might just be fortuitous since it's the second before last layer that correlates better with cross-lingual transfer for this task.

D Detailed results for alignment drop

Tab. 10 contains the detailed results when measuring the relative drop in strong alignment after fine-tuning. This is a detailed version of Tab. 3 in Section 5, with standard deviation measured over 5 different seeds for model initialization and data shuffling for fine-tuning, and all tested languages. This confirms that the observed increases and decreases in alignment are significant. It also seems to show that alignment for distant languages (en-ar, en-zh) is more affected by fine-tuning than other pairs.

E Breaking down correlation by models and layers

Tab. 12 shows a breakdown of the correlation between strong alignment and CTL across layers and models. These results tend to show that smaller models (distilmBERT and mBERT) have a better correlation at the last layer than larger models. It is also interesting to note that several correlation values are identical for alignment before fine-tuning, this might be explained by the fact that the seed of fine-tuning has unsurprisingly no effect on alignment measured before fine-tuning and by the possi-

task	model	en-ar	en-es	en-fr	en-ru	en-zh
POS	distilmBERT	-0.74$_{\pm 0.11}$	-0.86$_{\pm 0.04}$	-0.87$_{\pm 0.03}$	-0.87$_{\pm 0.01}$	-0.96$_{\pm 0.04}$
	mBERT	-0.90$_{\pm 0.04}$	-0.86$_{\pm 0.04}$	-0.93$_{\pm 0.02}$	-0.95$_{\pm 0.01}$	-0.96$_{\pm 0.02}$
	XLM-R Base	-0.43$_{\pm 0.18}$	-0.46$_{\pm 0.10}$	-0.46$_{\pm 0.17}$	-0.70$_{\pm 0.05}$	0.69$_{\pm 0.40}$
	XLM-R Large	-0.30$_{\pm 0.17}$	0.23$_{\pm 0.28}$	0.44$_{\pm 0.30}$	-0.44$_{\pm 0.14}$	0.26$_{\pm 0.25}$
NER	distilmBERT	0.00$_{\pm 0.37}$	-0.61$_{\pm 0.05}$	-0.60$_{\pm 0.05}$	-0.33$_{\pm 0.09}$	0.00$_{\pm 0.22}$
	mBERT	-0.28$_{\pm 0.19}$	-0.36$_{\pm 0.12}$	-0.49$_{\pm 0.11}$	-0.27$_{\pm 0.20}$	-0.25$_{\pm 0.13}$
	XLM-R Base	5.88$_{\pm 2.69}$	0.22$_{\pm 0.29}$	0.62$_{\pm 0.47}$	1.32$_{\pm 0.50}$	21.99$_{\pm 6.44}$
	XLM-R Large	16.34$_{\pm 8.76}$	2.22$_{\pm 0.44}$	3.17$_{\pm 0.89}$	3.10$_{\pm 0.72}$	12.67$_{\pm 3.97}$
NLI	distilmBERT	5.49$_{\pm 0.69}$	0.30$_{\pm 0.11}$	0.88$_{\pm 0.14}$	2.28$_{\pm 0.36}$	9.78$_{\pm 0.90}$
	mBERT	5.65$_{\pm 1.45}$	0.99$_{\pm 0.70}$	1.08$_{\pm 0.63}$	1.45$_{\pm 0.63}$	5.85$_{\pm 0.62}$
	XLM-R Base	11.17$_{\pm 0.95}$	1.01$_{\pm 0.12}$	1.55$_{\pm 0.28}$	2.67$_{\pm 0.24}$	27.58$_{\pm 3.33}$
	XLM-R Large	25.36$_{\pm 10.33}$	1.78$_{\pm 0.77}$	2.96$_{\pm 1.18}$	2.99$_{\pm 1.15}$	13.57$_{\pm 5.86}$

Table 10: Relative variation of strong alignment at the last layer before and after fine-tuning for different fine-tuning tasks. "$_\pm$" indicates standard deviation.

	before	after
last	0.89 (0.64 - 0.96)	0.83 (0.68 - 0.94)
-1	0.79 (0.63 - 0.89)	0.79 (0.64 - 0.88)
-2	0.79 (0.65 - 0.89)	0.78 (0.65 - 0.89)
-3	0.79 (0.64 - 0.89)	0.82 (0.69 - 0.91)
-4	0.79 (0.65 - 0.89)	0.76 (0.62 - 0.86)
-5	0.79 (0.64 - 0.89)	0.79 (0.64 - 0.91)
-6	0.79 (0.66 - 0.90)	0.77 (0.62 - 0.87)

Table 11: Correlation between strong English-target alignment and CTL from English to target language for the POS-tagging task, with 95% confidence intervals

bility that alignment measured at one layer might be almost perfectly correlated with alignment at another, especially when the correlation is measured across few languages.

However, drawing any conclusion from those figures might be irrelevant. By breaking down results by model, we measure correlation only from 25 samples, with five languages and five seeds. Furthermore, those latter seeds have no effect on alignment measured before. Tab. 11 shows a focus on distilmBERT for the same results with confidence intervals obtained with BCA bootstrapping. It demonstrates that the measured correlation is not precise enough to draw any conclusion on which layer has an alignment that is better correlated with CTL, or to determine whether alignment before or after fine-tuning is more relevant to CTL abilities. As a matter of fact, the results are so inconclusive that almost all correlation values in Tab. 12 lie in any of the confidence intervals in Tab. 11.

F Detailed results of the controlled experiment

This section provides detailed results of realignment methods for POS-tagging and NER, for all tested models, languages, translation datasets, and methods of extraction for realignment data. It also contains results for XNLI, for which only one translation dataset (opus100) and one extraction method (bilingual dictionaries) were tested. Results are shown on Tab. 13 (POS, opus100), 14 (POS, multi-UN), 15 (NER, opus100), 16 (NER, multi-UN), 17 (NLI, opus100).

A light gray cell indicates that the realignment method obtained an average score that is closer to the baseline with the same model than the standard deviation of the said baseline. A dark gray cell indicates that the realignment method provokes a decrease w.r.t. the baseline that is bigger than the standard deviation.

Those detailed results emphasize on the conclusions of Section 6. Using bilingual dictionaries seems to provide significant improvements more often than other methods to extract realignment pairs of words. This is particularly visible for the POS-tagging tasks, where realigning with a bilingual dictionary, with joint or sequential realignment, provides the best results. For the NER task, this is less visible, but we've already seen that, on average, bilingual dictionaries give better results (Tab. 5).

The detailed results also confirm that for smaller models and certain tasks like POS-tagging, realignment methods work better. Realignment methods for POS-tagging on distilmBERT bring significant

	distilmBERT		mBERT		XLM-R Base		XLM-R Large	
	before	after	before	after	before	after	before	after
last	0.89	0.83	0.89	0.82	0.59	0.67	0.75	0.78
-1	0.79	0.79	0.80	0.88	0.59	0.68	0.75	0.75
-2	0.79	0.78	0.80	0.84	0.59	0.68	0.75	0.76
-3	0.79	0.82	0.89	0.86	0.59	0.69	0.75	0.76
-4	0.79	0.76	0.89	0.83	0.59	0.68	0.75	0.74
-5	0.79	0.79	0.80	0.81	0.59	0.65	0.65	0.72
-6	0.79	0.77	0.80	0.80	0.69	0.64	0.65	0.70
-7	-	-	0.80	0.77	0.69	0.66	0.65	0.65
-8	-	-	0.80	0.77	0.69	0.66	0.65	0.62
-9	-	-	0.80	0.77	0.69	0.66	0.65	0.65
-10	-	-	0.80	0.79	0.69	0.68	0.65	0.69
-11	-	-	0.80	0.80	0.69	0.66	0.65	0.67
-12	-	-	0.89	0.89	0.69	0.68	0.65	0.67
-13	-	-	-	-	-	-	0.75	0.68
-14	-	-	-	-	-	-	0.75	0.76
-15	-	-	-	-	-	-	0.75	0.76
-16	-	-	-	-	-	-	0.75	0.76
-17	-	-	-	-	-	-	0.75	0.71
-18	-	-	-	-	-	-	0.75	0.72
-19	-	-	-	-	-	-	0.75	0.72
-20	-	-	-	-	-	-	0.75	0.73
-21	-	-	-	-	-	-	0.75	0.70
-22	-	-	-	-	-	-	0.75	0.70
-23	-	-	-	-	-	-	0.75	0.71
-24	-	-	-	-	-	-	0.75	0.75

Table 12: Correlation between strong English-target alignment and CTL from English to target language for the POS-tagging task. -i indicate depth of the model, with -1 being the second-before-last layer, and -2 the third-before-last, etc...

improvement for all languages. When using a bilingual dictionary, it also brings a systematically significant improvement over the baseline for mBERT on POS-tagging. For NER, the improvement is less often significant, but realignment methods still obtain some significant improvements for some languages like Arabic. For NLI, the only model on which there are some significant improvements for some languages is distilmBERT.

Using a supposedly higher quality translation dataset like multi-UN does not provide improvement over using opus100, which is said to be better reflecting the average quality of translation datasets (Wu and Dredze, 2020b). It might even seem that using multi-UN provide slightly worse results than opus100. There are more cases of unsignificant increase of results for multi-UN for POS-tagging and NER and also more cases of apparently significant degradation of results with respect to the baseline. This might be explained by the fact that multi-UN is a corpus obtained from translation of documents in the United Nations, which might lack diversity in their content.

Finally, we observe that realignment methods, at least with the small number of realignment steps we performed here, do not impact the evaluation on the fine-tuning language (English). Indeed, even if they sometimes provoke a decrease, namely on POS-tagging, this decrease is small, rarely of more than 0.1 points.

	en	ar	es	fr	ru	zh
distilmBERT	***96.1***$_{\pm0.1}$	*51.0*$_{\pm1.3}$	*84.1*$_{\pm0.8}$	*85.3*$_{\pm0.2}$	*81.2*$_{\pm0.7}$	*64.1*$_{\pm1.5}$
+ before fastalign	96.1$_{\pm0.0}$	63.4$_{\pm0.5}$	85.6$_{\pm0.1}$	86.5$_{\pm0.1}$	83.7$_{\pm0.5}$	66.3$_{\pm0.5}$
+ before awesome	96.1$_{\pm0.1}$	63.3$_{\pm0.9}$	85.4$_{\pm0.2}$	86.3$_{\pm0.1}$	82.9$_{\pm0.4}$	66.1$_{\pm0.5}$
+ before dico	96.1$_{\pm0.1}$	65.5$_{\pm0.5}$	85.8$_{\pm0.2}$	**86.5**$_{\pm0.2}$	**84.7**$_{\pm0.3}$	**67.4**$_{\pm0.7}$
+ joint fastalign	96.0$_{\pm0.1}$	62.9$_{\pm0.9}$	85.5$_{\pm0.2}$	86.2$_{\pm0.2}$	82.0$_{\pm0.5}$	65.0$_{\pm0.6}$
+ joint awesome	96.1$_{\pm0.1}$	63.4$_{\pm0.3}$	85.4$_{\pm0.1}$	86.4$_{\pm0.1}$	82.1$_{\pm0.5}$	64.8$_{\pm0.5}$
+ joint dico	96.1$_{\pm0.0}$	**66.8**$_{\pm0.6}$	**85.8**$_{\pm0.2}$	86.5$_{\pm0.2}$	84.1$_{\pm0.6}$	66.4$_{\pm0.7}$
mBERT	***96.7***$_{\pm0.0}$	*51.7*$_{\pm1.0}$	*85.6*$_{\pm0.3}$	*86.0*$_{\pm0.5}$	*82.1*$_{\pm0.7}$	*66.0*$_{\pm0.8}$
+ before fastalign	96.6$_{\pm0.1}$	64.2$_{\pm1.2}$	85.8$_{\pm0.2}$	86.5$_{\pm0.3}$	83.9$_{\pm0.8}$	67.3$_{\pm0.9}$
+ before awesome	96.6$_{\pm0.1}$	64.0$_{\pm1.2}$	85.9$_{\pm0.3}$	86.5$_{\pm0.4}$	83.4$_{\pm1.0}$	66.7$_{\pm0.9}$
+ before dico	96.6$_{\pm0.1}$	65.1$_{\pm0.6}$	**86.2**$_{\pm0.2}$	**86.9**$_{\pm0.3}$	**84.4**$_{\pm0.3}$	**68.3**$_{\pm0.6}$
+ joint fastalign	96.6$_{\pm0.0}$	62.8$_{\pm1.7}$	85.3$_{\pm0.3}$	86.5$_{\pm0.3}$	81.4$_{\pm0.4}$	65.2$_{\pm0.4}$
+ joint awesome	96.6$_{\pm0.1}$	63.3$_{\pm1.3}$	85.3$_{\pm0.2}$	86.4$_{\pm0.5}$	81.6$_{\pm0.5}$	65.7$_{\pm0.7}$
+ joint dico	96.6$_{\pm0.1}$	**66.5**$_{\pm0.8}$	86.1$_{\pm0.2}$	86.9$_{\pm0.3}$	83.9$_{\pm0.6}$	68.1$_{\pm0.8}$
XLM-R base	*95.9*$_{\pm0.1}$	*62.5*$_{\pm1.3}$	*86.6*$_{\pm0.3}$	*86.9*$_{\pm0.1}$	***86.9***$_{\pm0.6}$	*70.9*$_{\pm0.6}$
+ before fastalign	95.9$_{\pm0.1}$	64.2$_{\pm0.7}$	86.7$_{\pm0.1}$	87.3$_{\pm0.1}$	86.2$_{\pm0.6}$	68.5$_{\pm0.7}$
+ before awesome	**96.0**$_{\pm0.1}$	64.9$_{\pm1.5}$	86.8$_{\pm0.1}$	87.2$_{\pm0.1}$	86.3$_{\pm0.7}$	68.0$_{\pm0.7}$
+ before dico	96.0$_{\pm0.1}$	**67.3**$_{\pm1.5}$	**86.9**$_{\pm0.2}$	**87.3**$_{\pm0.1}$	86.8$_{\pm0.7}$	**71.2**$_{\pm0.6}$
+ joint fastalign	95.9$_{\pm0.1}$	63.5$_{\pm0.4}$	86.0$_{\pm0.2}$	86.6$_{\pm0.2}$	84.6$_{\pm0.2}$	69.1$_{\pm0.5}$
+ joint awesome	96.0$_{\pm0.1}$	63.1$_{\pm0.7}$	85.8$_{\pm0.1}$	86.4$_{\pm0.1}$	84.5$_{\pm0.3}$	69.2$_{\pm0.2}$
+ joint dico	95.9$_{\pm0.1}$	66.6$_{\pm0.4}$	86.6$_{\pm0.2}$	87.2$_{\pm0.1}$	86.0$_{\pm0.3}$	70.6$_{\pm0.2}$
XLM-R large	***97.7***$_{\pm0.0}$	***65.1***$_{\pm0.6}$	***87.0***$_{\pm0.6}$	***87.5***$_{\pm0.6}$	***87.0***$_{\pm0.9}$	***71.5***$_{\pm0.2}$

Table 13: Controlled experiment with realignment for POS-tagging with opus100 translation dataset.

	en	ar	es	fr	ru	zh
distilmBERT	*96.1*$_{\pm0.1}$	*51.0*$_{\pm1.3}$	*84.1*$_{\pm0.8}$	*85.3*$_{\pm0.2}$	*81.2*$_{\pm0.7}$	*64.1*$_{\pm1.5}$
+ before fastalign	96.0$_{\pm0.1}$	61.8$_{\pm0.6}$	85.2$_{\pm0.2}$	86.1$_{\pm0.2}$	82.0$_{\pm0.4}$	65.7$_{\pm0.7}$
+ before awesome	96.0$_{\pm0.1}$	62.1$_{\pm0.4}$	85.3$_{\pm0.2}$	86.1$_{\pm0.4}$	82.2$_{\pm0.5}$	65.3$_{\pm0.5}$
+ before dico	96.1$_{\pm0.0}$	64.1$_{\pm0.9}$	**85.8**$_{\pm0.1}$	86.5$_{\pm0.1}$	**84.6**$_{\pm0.6}$	**67.4**$_{\pm1.0}$
+ joint fastalign	96.1$_{\pm0.1}$	62.8$_{\pm0.6}$	85.2$_{\pm0.1}$	86.1$_{\pm0.1}$	81.2$_{\pm0.4}$	64.7$_{\pm0.6}$
+ joint awesome	**96.1**$_{\pm0.1}$	61.8$_{\pm0.8}$	85.2$_{\pm0.2}$	86.0$_{\pm0.2}$	81.2$_{\pm0.4}$	64.6$_{\pm0.5}$
+ joint dico	96.1$_{\pm0.0}$	**65.5**$_{\pm0.4}$	85.8$_{\pm0.2}$	**86.7**$_{\pm0.1}$	83.7$_{\pm0.6}$	65.9$_{\pm0.8}$
mBERT	*96.7*$_{\pm0.0}$	*51.7*$_{\pm1.0}$	*85.6*$_{\pm0.3}$	*86.0*$_{\pm0.5}$	*82.1*$_{\pm0.7}$	*66.0*$_{\pm0.8}$
+ before fastalign	96.6$_{\pm0.0}$	63.1$_{\pm0.6}$	85.6$_{\pm0.1}$	86.4$_{\pm0.2}$	82.7$_{\pm0.6}$	66.9$_{\pm0.9}$
+ before awesome	**96.7**$_{\pm0.1}$	61.8$_{\pm1.0}$	85.6$_{\pm0.3}$	86.5$_{\pm0.3}$	82.1$_{\pm0.8}$	66.6$_{\pm0.8}$
+ before dico	96.6$_{\pm0.1}$	64.2$_{\pm1.4}$	**86.0**$_{\pm0.3}$	86.9$_{\pm0.4}$	**84.1**$_{\pm0.8}$	**69.0**$_{\pm0.8}$
+ joint fastalign	96.7$_{\pm0.0}$	62.6$_{\pm1.0}$	85.4$_{\pm0.3}$	86.3$_{\pm0.5}$	80.7$_{\pm0.8}$	65.0$_{\pm0.6}$
+ joint awesome	96.7$_{\pm0.0}$	61.9$_{\pm0.8}$	85.2$_{\pm0.3}$	86.1$_{\pm0.3}$	80.9$_{\pm0.4}$	64.9$_{\pm0.6}$
+ joint dico	96.6$_{\pm0.1}$	**64.5**$_{\pm1.4}$	86.0$_{\pm0.4}$	**86.9**$_{\pm0.5}$	83.9$_{\pm0.9}$	67.3$_{\pm0.7}$
XLM-R base	*95.9*$_{\pm0.1}$	*62.5*$_{\pm1.3}$	*86.6*$_{\pm0.3}$	*86.9*$_{\pm0.1}$	***86.9***$_{\pm0.6}$	*70.9*$_{\pm0.6}$
+ before fastalign	95.9$_{\pm0.1}$	64.0$_{\pm1.0}$	86.3$_{\pm0.1}$	87.0$_{\pm0.3}$	85.8$_{\pm0.7}$	68.6$_{\pm0.9}$
+ before awesome	**96.0**$_{\pm0.1}$	64.7$_{\pm0.9}$	86.4$_{\pm0.1}$	86.8$_{\pm0.2}$	85.8$_{\pm0.4}$	66.8$_{\pm0.0}$
+ before dico	96.0$_{\pm0.1}$	**66.5**$_{\pm1.2}$	**86.8**$_{\pm0.2}$	**87.2**$_{\pm0.2}$	86.3$_{\pm0.6}$	**71.1**$_{\pm0.6}$
+ joint fastalign	95.9$_{\pm0.1}$	62.5$_{\pm1.1}$	85.7$_{\pm0.2}$	86.3$_{\pm0.2}$	84.1$_{\pm0.4}$	69.1$_{\pm0.3}$
+ joint awesome	95.9$_{\pm0.1}$	62.1$_{\pm0.9}$	85.5$_{\pm0.1}$	86.2$_{\pm0.2}$	83.9$_{\pm0.3}$	69.0$_{\pm0.4}$
+ joint dico	95.9$_{\pm0.1}$	65.8$_{\pm0.7}$	86.4$_{\pm0.3}$	87.1$_{\pm0.1}$	85.3$_{\pm0.6}$	70.5$_{\pm0.2}$
XLM-R large	***97.7***$_{\pm0.0}$	***65.1***$_{\pm0.6}$	***87.0***$_{\pm0.6}$	***87.5***$_{\pm0.6}$	***87.0***$_{\pm0.9}$	***71.5***$_{\pm0.2}$

Table 14: Controlled experiment with realignment for POS-tagging with multiUN translation dataset.

	en	ar	es	fr	ru	zh
distilmBERT	*82.9*$_{\pm0.4}$	*34.5*$_{\pm1.6}$	*69.2*$_{\pm3.1}$	*76.1*$_{\pm0.7}$	*60.2*$_{\pm0.9}$	*46.8*$_{\pm1.9}$
+ before fastalign	82.9$_{\pm0.3}$	39.1$_{\pm2.0}$	70.1$_{\pm2.7}$	75.9$_{\pm0.3}$	60.1$_{\pm0.8}$	46.5$_{\pm2.5}$
+ before awesome	82.7$_{\pm0.3}$	39.3$_{\pm3.7}$	72.5$_{\pm2.5}$	75.6$_{\pm0.5}$	60.3$_{\pm0.3}$	**49.7**$_{\pm1.1}$
+ before dico	82.9$_{\pm0.4}$	41.6$_{\pm1.7}$	67.9$_{\pm2.7}$	76.4$_{\pm0.9}$	60.3$_{\pm1.2}$	48.3$_{\pm1.6}$
+ joint fastalign	83.0$_{\pm0.2}$	41.5$_{\pm3.1}$	**73.5**$_{\pm2.1}$	76.6$_{\pm0.7}$	**61.4**$_{\pm0.8}$	48.3$_{\pm1.5}$
+ joint awesome	**83.1**$_{\pm0.1}$	41.4$_{\pm0.4}$	72.3$_{\pm0.1}$	**77.6**$_{\pm0.7}$	60.9$_{\pm0.4}$	49.5$_{\pm0.4}$
+ joint dico	83.0$_{\pm0.5}$	**42.2**$_{\pm2.7}$	69.5$_{\pm2.3}$	76.6$_{\pm1.0}$	61.2$_{\pm0.8}$	48.8$_{\pm1.7}$
mBERT	*84.4*$_{\pm0.4}$	*40.7*$_{\pm2.9}$	*74.3*$_{\pm1.4}$	*79.9*$_{\pm1.3}$	*63.9*$_{\pm2.0}$	*52.1*$_{\pm1.7}$
+ before fastalign	84.3$_{\pm0.4}$	42.0$_{\pm2.9}$	70.5$_{\pm3.0}$	79.3$_{\pm0.6}$	**65.5**$_{\pm1.6}$	51.7$_{\pm1.1}$
+ before awesome	**84.8**$_{\pm0.2}$	40.2$_{\pm2.6}$	72.3$_{\pm2.8}$	79.4$_{\pm0.7}$	63.0$_{\pm1.8}$	51.2$_{\pm0.6}$
+ before dico	84.3$_{\pm0.6}$	42.1$_{\pm1.7}$	73.4$_{\pm2.8}$	80.1$_{\pm1.4}$	64.9$_{\pm1.7}$	52.8$_{\pm1.2}$
+ joint fastalign	84.3$_{\pm0.3}$	42.7$_{\pm1.9}$	75.6$_{\pm2.0}$	80.5$_{\pm1.0}$	65.4$_{\pm1.5}$	54.3$_{\pm1.2}$
+ joint awesome	84.1$_{\pm0.4}$	44.2$_{\pm2.2}$	75.6$_{\pm1.6}$	80.2$_{\pm0.2}$	64.8$_{\pm2.4}$	54.6$_{\pm1.1}$
+ joint dico	84.2$_{\pm0.3}$	**46.0**$_{\pm3.2}$	**76.6**$_{\pm1.9}$	**81.1**$_{\pm0.9}$	65.5$_{\pm0.9}$	**54.9**$_{\pm0.9}$
XLM-R base	*80.0*$_{\pm0.3}$	*46.4*$_{\pm2.7}$	*71.8*$_{\pm3.7}$	*75.0*$_{\pm1.4}$	*61.6*$_{\pm0.8}$	*47.4*$_{\pm2.1}$
+ before fastalign	**80.2**$_{\pm0.4}$	51.5$_{\pm3.1}$	71.7$_{\pm1.4}$	75.9$_{\pm1.0}$	62.1$_{\pm1.2}$	45.7$_{\pm1.3}$
+ before awesome	80.1$_{\pm0.3}$	52.2$_{\pm3.4}$	74.2$_{\pm1.3}$	76.1$_{\pm0.8}$	61.0$_{\pm1.7}$	46.6$_{\pm1.0}$
+ before dico	80.0$_{\pm0.2}$	**55.8**$_{\pm3.6}$	**76.9**$_{\pm1.6}$	**77.3**$_{\pm0.7}$	62.0$_{\pm0.4}$	47.5$_{\pm0.7}$
+ joint fastalign	79.8$_{\pm0.2}$	47.7$_{\pm5.0}$	74.2$_{\pm1.2}$	75.6$_{\pm0.7}$	63.0$_{\pm0.8}$	50.2$_{\pm1.5}$
+ joint awesome	79.7$_{\pm0.3}$	47.6$_{\pm3.1}$	73.9$_{\pm1.2}$	75.4$_{\pm0.4}$	63.5$_{\pm0.8}$	**51.2**$_{\pm1.1}$
+ joint dico	79.9$_{\pm0.3}$	50.3$_{\pm3.2}$	75.2$_{\pm1.1}$	75.9$_{\pm0.8}$	**63.6**$_{\pm0.6}$	50.1$_{\pm1.3}$
XLM-R large	***83.8***$_{\pm1.0}$	***45.1***$_{\pm1.4}$	***75.6***$_{\pm3.7}$	***80.7***$_{\pm0.8}$	***70.5***$_{\pm3.4}$	***53.0***$_{\pm2.1}$

Table 15: Controlled experiment with realignment for NER with opus100 translation dataset.

	same	ar	es	fr	ru	zh
distilmBERT	*82.9*$_{\pm0.4}$	*34.5*$_{\pm1.6}$	*69.2*$_{\pm3.1}$	*76.1*$_{\pm0.7}$	*60.2*$_{\pm0.9}$	*46.8*$_{\pm1.9}$
+ before fastalign	82.9$_{\pm0.3}$	37.8$_{\pm0.6}$	71.1$_{\pm3.8}$	76.2$_{\pm1.3}$	59.8$_{\pm1.7}$	46.9$_{\pm2.0}$
+ before awesome	83.0$_{\pm0.4}$	39.4$_{\pm1.1}$	**71.7**$_{\pm3.6}$	75.7$_{\pm0.3}$	61.2$_{\pm1.7}$	46.3$_{\pm1.3}$
+ before dico	**83.0**$_{\pm0.1}$	**43.3**$_{\pm2.9}$	69.0$_{\pm3.5}$	**77.6**$_{\pm1.0}$	59.9$_{\pm0.9}$	47.1$_{\pm1.5}$
+ joint fastalign	83.0$_{\pm0.2}$	39.5$_{\pm0.4}$	70.5$_{\pm2.3}$	76.6$_{\pm0.5}$	61.3$_{\pm1.0}$	48.4$_{\pm1.7}$
+ joint awesome	82.9$_{\pm0.3}$	38.8$_{\pm1.8}$	69.6$_{\pm1.0}$	76.7$_{\pm0.8}$	60.6$_{\pm1.4}$	**49.2**$_{\pm1.1}$
+ joint dico	83.0$_{\pm0.3}$	41.9$_{\pm1.5}$	71.3$_{\pm1.7}$	77.5$_{\pm1.5}$	**61.9**$_{\pm1.4}$	48.4$_{\pm1.1}$
mBERT	*84.4*$_{\pm0.4}$	*40.7*$_{\pm2.9}$	*74.3*$_{\pm1.4}$	*79.9*$_{\pm1.3}$	*63.9*$_{\pm2.0}$	*52.1*$_{\pm1.7}$
+ before fastalign	84.3$_{\pm0.4}$	**46.2**$_{\pm3.6}$	69.4$_{\pm4.2}$	78.4$_{\pm0.8}$	64.3$_{\pm1.1}$	51.2$_{\pm1.9}$
+ before dico	**84.6**$_{\pm0.4}$	42.5$_{\pm2.0}$	71.9$_{\pm2.6}$	79.9$_{\pm0.8}$	64.0$_{\pm1.1}$	52.1$_{\pm1.1}$
+ joint fastalign	84.1$_{\pm0.3}$	46.0$_{\pm1.0}$	74.4$_{\pm2.6}$	**80.7**$_{\pm1.0}$	**66.6**$_{\pm2.4}$	**54.8**$_{\pm0.4}$
+ joint awesome	84.1$_{\pm0.5}$	42.2$_{\pm2.2}$	**74.9**$_{\pm0.5}$	80.3$_{\pm0.6}$	65.5$_{\pm2.5}$	54.5$_{\pm0.9}$
+ joint dico	84.3$_{\pm0.4}$	43.6$_{\pm3.2}$	74.5$_{\pm1.2}$	80.4$_{\pm0.8}$	65.6$_{\pm3.3}$	53.4$_{\pm1.5}$
XLM-R base	*80.0*$_{\pm0.3}$	*46.4*$_{\pm2.7}$	*71.8*$_{\pm3.7}$	*75.0*$_{\pm1.4}$	*61.6*$_{\pm0.8}$	*47.4*$_{\pm2.1}$
+ before fastalign	**80.2**$_{\pm0.3}$	53.8$_{\pm3.4}$	72.0$_{\pm2.8}$	76.0$_{\pm1.4}$	61.8$_{\pm1.5}$	45.8$_{\pm1.3}$
+ before awesome	80.2$_{\pm0.3}$	**54.8**$_{\pm1.3}$	70.4$_{\pm2.0}$	76.5$_{\pm1.4}$	61.9$_{\pm0.0}$	45.5$_{\pm0.7}$
+ before dico	80.0$_{\pm0.2}$	54.1$_{\pm1.7}$	**76.1**$_{\pm1.4}$	**76.5**$_{\pm0.7}$	61.8$_{\pm1.0}$	47.6$_{\pm1.4}$
+ joint fastalign	79.8$_{\pm0.2}$	46.8$_{\pm3.0}$	73.5$_{\pm2.4}$	75.6$_{\pm1.0}$	61.8$_{\pm1.5}$	50.0$_{\pm1.8}$
+ joint awesome	79.8$_{\pm0.3}$	49.0$_{\pm3.1}$	75.8$_{\pm1.7}$	76.1$_{\pm1.1}$	**63.0**$_{\pm1.5}$	**50.8**$_{\pm0.8}$
+ joint dico	79.8$_{\pm0.3}$	48.4$_{\pm3.0}$	74.3$_{\pm1.6}$	75.8$_{\pm0.8}$	62.7$_{\pm0.7}$	50.2$_{\pm0.1}$
XLM-R large	***83.8***$_{\pm1.0}$	***45.1***$_{\pm1.4}$	***75.6***$_{\pm3.7}$	***80.7***$_{\pm0.8}$	***70.5***$_{\pm3.4}$	***53.0***$_{\pm2.1}$

Table 16: Controlled experiment with realignment for NER with multiUN translation dataset.

	en	ar	es	fr	ru	zh
distilmBERT	*76.0*$_{\pm0.7}$	*58.2*$_{\pm1.4}$	*68.5*$_{\pm0.6}$	***68.7***$_{\pm0.6}$	*62.3*$_{\pm1.2}$	*63.4*$_{\pm0.9}$
+ before dico	**76.2**$_{\pm0.6}$	58.4$_{\pm1.2}$	69.2$_{\pm0.8}$	68.0$_{\pm0.8}$	62.6$_{\pm1.1}$	63.1$_{\pm0.9}$
+ joint dico	76.2$_{\pm0.7}$	**59.8**$_{\pm1.1}$	**69.2**$_{\pm1.0}$	68.6$_{\pm1.1}$	**63.0**$_{\pm1.0}$	**64.4**$_{\pm1.2}$
mBERT	***80.2***$_{\pm0.7}$	*65.2*$_{\pm1.2}$	*73.8*$_{\pm0.8}$	***72.9***$_{\pm0.7}$	*68.3*$_{\pm1.2}$	*68.5*$_{\pm1.3}$
+ before dico	79.0$_{\pm0.6}$	63.2$_{\pm0.9}$	72.9$_{\pm0.8}$	71.7$_{\pm0.5}$	66.9$_{\pm0.9}$	68.7$_{\pm0.7}$
+ joint dico	79.9$_{\pm0.9}$	**65.6**$_{\pm0.9}$	**73.8**$_{\pm1.3}$	72.5$_{\pm1.2}$	**68.8**$_{\pm1.3}$	**69.0**$_{\pm0.8}$
XLM-R base	*82.8*$_{\pm1.6}$	*70.1*$_{\pm1.4}$	*77.4*$_{\pm1.6}$	*76.5*$_{\pm1.3}$	*74.2*$_{\pm1.3}$	*71.7*$_{\pm1.3}$
+ before dico	81.2$_{\pm2.4}$	68.4$_{\pm2.9}$	76.0$_{\pm2.2}$	74.9$_{\pm1.9}$	72.8$_{\pm2.4}$	71.4$_{\pm2.5}$
+ joint dico	**83.7**$_{\pm0.7}$	**70.8**$_{\pm1.4}$	**78.0**$_{\pm0.9}$	**76.7**$_{\pm1.0}$	**74.6**$_{\pm1.3}$	**72.7**$_{\pm1.6}$
XLM-R large	***87.9***$_{\pm0.7}$	***77.5***$_{\pm1.3}$	***83.2***$_{\pm1.3}$	***81.9***$_{\pm1.2}$	***79.1***$_{\pm1.1}$	***78.2***$_{\pm1.3}$

Table 17: Controlled experiment with realignment for NLI with opus100 translation dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☒ A2. Did you discuss any potential risks of your work?
Our work does not introduce new methods. It is anlysing existing ones and trying to provide a better understanding of their inner working. Hence, we do not believe that our paper present a direct risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Introduction and abstract.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Appendix A.1

☑ B1. Did you cite the creators of artifacts you used?
Appendix A.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All artifacts used are under permissive licences or terms of use.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Provided it was specified, our use of existing artifact was consistent with intended use.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We did not collect data, and the datasets we used are already publicly available.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 8. Limitations

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A.

C ☑ Did you run computational experiments?

Section 4 and 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.6 and Table "Number of parameters" in Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Overlap with standard deviation is indicated on all tables, error bars are provided on all graphs where it is relevant. Appendices C, D, E, F provide detailed results with confidence intervals and standard deviations.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.