



HAL
open science

Multilingual Transformer Encoders: a Word-Level Task-Agnostic Evaluation

Félix Gaschi, François Plesse, Parisa Rastin, Yannick Toussaint

► **To cite this version:**

Félix Gaschi, François Plesse, Parisa Rastin, Yannick Toussaint. Multilingual Transformer Encoders: a Word-Level Task-Agnostic Evaluation. WCCI2022 - IEEE World Congress on Computational Intelligence, Jul 2022, Padoue, Italy. hal-03723760

HAL Id: hal-03723760

<https://hal.science/hal-03723760>

Submitted on 15 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multilingual Transformer Encoders: a Word-Level Task-Agnostic Evaluation

Félix Gaschi
team ORPAILLEUR
LORIA

Vandœuvre-lès-Nancy, France
felix.gaschi@loria.fr

François Plesse
R&D dpt.
Posos

Paris, France
francois@posos.fr

Parisa Rastin
team ABC
LORIA

Vandœuvre-lès-Nancy, France
parisa.rastin@loria.fr

Yannick Toussaint
team ORPAILLEUR
LORIA

Vandœuvre-lès-Nancy, France
yannick.toussaint@loria.fr

Abstract—Some Transformer-based models can perform cross-lingual transfer learning: those models can be trained on a specific task in one language and give relatively good results on the same task in another language, despite having been pre-trained on monolingual tasks only. But, there is no consensus yet on whether those transformer-based models learn universal patterns across languages. We propose a word-level task-agnostic method to evaluate the alignment of contextualized representations built by such models. We show that our method provides more accurate translated word pairs than previous methods to evaluate word-level alignment. And our results show that some inner layers of multilingual Transformer-based models outperform other explicitly aligned representations, and even more so according to a stricter definition of multilingual alignment.

Index Terms—Deep learning, Natural language processing, Text mining

I. INTRODUCTION

Building aligned multilingual representations is namely useful for cross-lingual document retrieval. For example, a query written in German, might be answered with documents in German but also in other languages, like scientific papers in English. Aligned multilingual representations would allow to embed all documents as well as the query in a shared space, allowing document retrieval by nearest-neighbor search. Good word-level alignment could lead to more fine-grained cross-lingual information retrieval, like passage retrieval, allowing for extracting relevant passages from relevant documents in multiple languages.

This paper provides an evaluation of the word-level multilingual alignment of contextualized representations produced by multilingual models, particularly by multilingual Transformer encoders [1]. This evaluation framework does not rely on any specific probing task and directly compares word representations instead of sentence-level ones [2], [3]. The proposed method relies on a bilingual dictionary instead of a probabilistic tool like FastAlign [4], as in [5], to retrieve more accurate word pairs in parallel texts. It aims at measuring directly how well word representations from different languages are aligned and at producing a comparative analysis of existing models.

To build accurate language representations, models based on the Transformer architecture were introduced a few years ago with BERT [6] and now play a key role in natural language processing (NLP) research. Those models are first

pre-trained in an unsupervised manner on large text datasets and can then be fine-tuned on a downstream task like sentence classification, named entity recognition (NER), or extractive question answering. BERT is also used as a building block for document retrieval models like ColBERT [7].

NLP research is mainly led around the English language [8]. BERT, and many of its proposed variations like RoBERTa [9], SciBERT [10], or BioBERT [11], are pre-trained and evaluated solely in English. Some BERT-like models were proposed in other languages, like CamemBERT [12] and FlauBERT [13] in French, FinBERT [14] (Finnish), AfriBERT [15] (Afrikaans), etc. They bring the advances of Transformer-based models to yet another language but they are still monolingual.

Multilingual Transformer-based models, like mBERT [6] or XLM-R [16], are the focus of our paper. They are pre-trained with the same objective as their monolingual counterparts but on several monolingual corpora in different languages.

Although they have never been explicitly trained on parallel text, those models were shown to be able to generalize well from one language to another. For instance, instead of fine-tuning and evaluating in the same language after pre-training, the models can be fine-tuned on an English Named Entity Recognition (NER) training dataset and evaluated on a French one and still obtain good results. This process was dubbed “Cross-lingual Transfer Learning” (CTL) [2].

Despite those cross-lingual generalization abilities, there is no consensus on whether those models exhibit aligned multilingual representations. By aligned representations, we mean that a word and its translation should be attributed similar representations, in the same way that words or sentences of similar meaning should be represented by similar representations in a monolingual embedding.

We want to directly evaluate the quality of the alignment in multilingual models because cross-lingual transfer learning is not necessarily correlated with multilingual alignment. Our contribution is (1) to propose a method for extracting pairs of translated words in context; (2) to show that the proposed method extracts more accurate pairs than others [5] which rely on word-alignment tools like FastAlign [4]; and (3) to reveal that the alignment produced by most Transformer-based multilingual models is competitive with other representations.

II. RELATED WORK

Building universal multilingual representations has been envisioned long before the introduction of the Transformer architecture with ideas like Chomsky’s “Universal Grammar” [17] and the “linguistic universals” of Greenberg [18]. But Transformer-based models are expected to build such representations “at scale”, according to [16], with the goal that low-resource languages might benefit from high-resource ones by sharing some features like common sub-words [2]. However, many works have focused on the ability to perform cross-lingual transfer learning [2], [16], [19] and less on the multilingual alignment of the produced representation [20].

A. Word Embedding Alignment

When word embeddings were introduced [21], it was then shown that two or more monolingual embeddings could be aligned effectively [22]. State-of-the-art alignment methods for word embeddings give impressive results on bilingual lexicon induction (BLI), a task where the translation of a word is retrieved with nearest-neighbor search [23]–[25]. In our experiments, a specific version of those aligned word embeddings will be used as a baseline: monolingual FastText embeddings [26] aligned with RCLS [24].

A multilingual word embedding provides a static representation of each word, meaning that a word will always be represented by the same vector regardless of the context, whereas the Transformer-based models which are the focus of our paper provide a contextualized representation of a word. The representation attributed to a word by such a deeper model varies according to the surrounding words.

B. Multilingual Language Models

Deeper multilingual language models were first proposed as a way to perform unsupervised machine translation [27]. Besides unsupervised translation models, LASER [28] is a multilingual sentence embedding based on RNNs and trained on several parallel corpora. MultiFiT [29] is a training procedure involving LASER and a monolingual model to perform a cross-lingual transfer. However, we will focus here on Transformer-based models which are pre-trained with fewer parallel texts and are not built specifically for sentence-level downstream tasks like LASER and MultiFit.

C. Multilingual Transformer Encoders

The Transformer architecture was initially proposed as an encoder-decoder architecture [1], for sequence-to-sequence tasks like translation. But BERT [6] uses only the encoder part of the architecture to set a new standard for tasks that require encoding the full sentence like sentence classification, named entity recognition, or extractive question answering. BERT is typically pre-trained to predict randomly masked-out words across a wide English corpus. A multilingual version of BERT, called mBERT, was proposed and is the same architecture pre-trained for the same task but on 104 monolingual corpora of distinct languages.

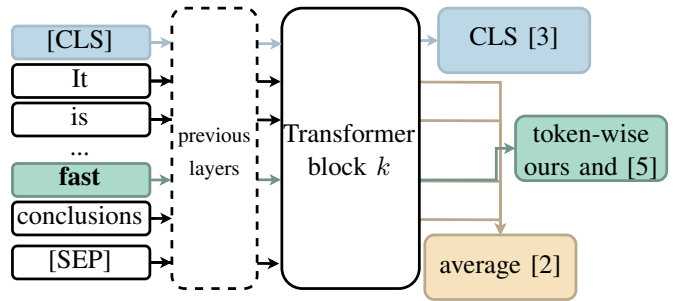


Fig. 1. Different representations extracted from multilingual Transformers: sentences representations [2], [3] or token representations, ours and [5].

Despite not being pre-trained explicitly on parallel text, mBERT was shown to have surprisingly good cross-lingual transfer abilities [2], [19]. For example, when mBERT is fine-tuned on a classification task in English, it gives competitive results when evaluated in French. This has also been shown to work between typologically different languages, although to a lesser extent than similar languages.

Multilingual language models (MLLMs) are often either framed as a way to perform unsupervised machine translation or either as a ground for generalizing the learning of downstream tasks across languages. In this paper, we wonder whether MLLMs can hold “universal” patterns across languages, or at least patterns that are common to a wide range of different languages, and produce aligned representations.

D. Absence of consensus about multilingual alignment

Despite the surprisingly efficient cross-lingual transfer of models like mBERT, there is no consensus on whether those multilingual models learn universal multilingual patterns, as a recent literature review states [20]. With some probing tasks, it is shown that we can learn syntactic trees, which represent the grammatical structure of a sentence, that are consistent across languages with heuristics applied to attention activation values [30]. But other probing tasks show that the same models retain language-specific information at every layer [19], [31].

Another approach for studying multilingual patterns learned by a model is to perform inferences on parallel text. Different approaches lead to opposite results [2], [3]. As shown in Fig. 1, [2] compares sentence representations obtained by averaging sub-word representations, whereas [3] uses the representation of the initial token [CLS] which is supposed to hold a representation of the sentence. The former found aligned representations, while the latter concluded that mBERT “is not an interlanguage”. We propose to compare word representations instead similarly to [5].

Indeed, even if the sentence-level alignment was guaranteed, does it necessarily mean that there is a word-level multilingual alignment in the representations produced by mBERT and others? the CLS representation is independent of the token representation. And if the same information was repeated in

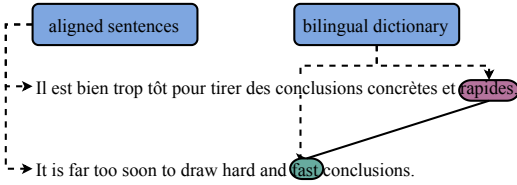


Fig. 2. Extracting a contextualized translated pair with a bilingual dictionary.

each token embedding, one could have an alignment of average sentence representation, but not of the single tokens.

It was shown that mBERT builds representations that enable the extraction of corresponding words in a pair of translated sentences with fair accuracy, and AWESOME [32] was proposed to fine-tune mBERT and improve this word-level matching. In another work [5], translated pairs of words were extracted from translated sentences with FastAlign [4]. The similarity between the contextualized representations built by mBERT for those words was compared with the similarity between representations of random words. They observed that the similarity for translated words was close to that of random pairs. We show that tools like FastAlign make too many alignment mistakes to be able to conclude about the quality of the multilingual alignment produced by models like mBERT. Our method relies on the extraction of translated words from translated sentences in a manner that is less prone to errors for evaluating multilingual models.

III. METHODOLOGY

A. Translated-in-context word pairs

In order to extract translated words from translated sentences with a minimized number of errors, we start from a bilingual dictionary. And instead of trying to extract all possible word pairs with a word-alignment tool like FastAlign, we extract only those that we are certain of.

Our method requires two datasets: a translation dataset containing gold standard translated pairs of sentences and a bilingual dictionary. The bilingual dictionary contains pairs of translated words, like the pair "rapide"- "fast" in our example Fig. 2. In a pair of translated sentences from the translation dataset, for each word from one sentence, every potential translated word indicated by the bilingual dictionary is collected from the other sentence. A pair of translated words is kept with its context if there is only one candidate for the translation.

The pairs obtained can be seen as a contextualized bilingual dictionary, where translated words go along with their context. We argue that we could not use pairs of words from the original bilingual dictionary directly because the models we want to evaluate are pre-trained on whole sentences and, since they have not been pre-trained on single words, they might not give relevant representations of words without context.

Working at the word level instead of the sentence level like [3] and [2] is chosen for several reasons. A good alignment of sentence representations does not necessarily guarantee a good alignment of word representations. Also, working at the

word level allows a more direct comparison with multilingual word embeddings, which can then be used as a baseline.

B. Inference on sentences

Once we have extracted translated-in-context pairs of words, they are passed through the model we want to evaluate to produce contextualized representations of the translated words.

To build the representation of each contextualized word, the whole sentences are passed to the evaluated model (typically mBERT) and the contextualized representations of the words from the extracted word pairs are kept. We can extract such representations for each stacked Transformer block (or layer). By convention, the representation of index 0 will be the one from the input un-contextualized embedding layer. For example, mBERT, which uses 12 Transformer blocks, will produce 13 representations for a word, the 0th one for the initial embedding and the 1st to 12th for the output of each Transformer block.

C. Nearest-neighbor retrieval task

For each layer of a Transformer-based model, a nearest-neighbor retrieval task is then performed on the produced representations. In a similar manner as [2], a fixed number N of representations of translated-in-context pairs is randomly sampled. Then, for each of those sampled pairs of representations (u_i, v_i) , the aim is to check that v_i is the nearest neighbor of u_i with respect to all other v_j . The final evaluation score is then the proportion of pairs for which the translation is the nearest neighbor. More formally our score is given by:

$$S_{\text{weak}}(U, V) = \frac{1}{N} \sum_{i=0}^N \mathbb{1} \left[s(u_i, v_i) > \max_{j \neq i} s(u_i, v_j) \right] \quad (1)$$

Where U and V are the sets of sampled u_i and v_i respectively, s is a similarity function or retrieval criterion and N the number of sampled pairs.

The retrieval criterion used is the cross-domain similarity local scaling (CSLS) [23]. It is a modified cosine similarity commonly used for retrieval in word embeddings [23]–[25] which takes into account the density around the compared representations by deducing from the initial similarity the averaged cosine similarity of the k nearest neighbor of each representation in the pair:

$$s(u_i, v_j) = 2 \cos(u_i, v_j) - \frac{1}{k} \sum_{v \in \mathcal{N}_V^k(u_i)} \cos(u_i, v) - \frac{1}{k} \sum_{u \in \mathcal{N}_U^k(v_j)} \cos(u, v_j) \quad (2)$$

c is the cosine similarity and $\mathcal{N}_V^k(u_i)$ the set of k nearest neighbors of u_i in V .

In our reported experiments, we used a similarity based on the cosine similarity and not on the l2-distance for two reasons: (1) the CSLS criterion is commonly used for evaluating multilingual word embeddings which is our baseline and (2) we

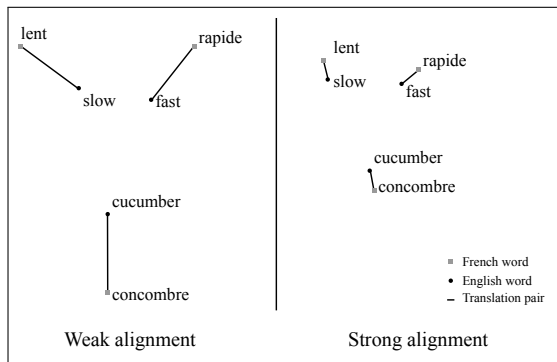


Fig. 3. Toy example illustrating weak and strong alignment.

noticed that l2-distance gives similar results for the evaluated models which can be explained by the Transformer architecture using a layer normalization at the end of each Transformer block. If the representations have roughly constant norms, then l2-distance is approximately proportional to cosine distance.

D. Evaluating strong alignment

With the retrieval score S_{weak} described in Equation 1, we are only evaluating “weak” alignment as described by [33]. Weak alignment is defined as the fact that the nearest neighbor in language L_1 of any item in a language L_2 is the most relevant item (i.e. the translation in our case). For example, as shown in Fig. 3, “fast” will be closer to “rapide” than “slow” or “cucumber” but “lent” or “concombre” can be even closer.

Strong alignment is defined as the fact that translated pairs are closer than irrelevant pairs, regardless of the language. In this case, “fast” should be closer to “rapide” than “slow” and “cucumber”, but also than “lent” and “concombre”. Strong alignment is useful for document retrieval. Indeed, if a document retrieval system is expected to answer a query with documents from different languages, it should be able to rank documents across languages.

We propose another retrieval score like in Eq. 1 which evaluates strong alignment by checking that for each pair of representations (u_i, v_i) the translation v_i is closer to u_i than any element of U instead of V :

$$S_{\text{strong}}(U, V) = \frac{1}{N} \sum_{i=0}^N \mathbb{1} \left[s(u_i, v_i) > \max_{j \neq i} s(u_i, u_j) \right] \quad (3)$$

The only change is in bold: v_j has been replaced by u_j . For instance, we want the representation of the French word “rapide” to be closer to its translation “fast” than any other English word (instead of French for S_{weak}).

IV. EXPERIMENTAL SETUP

We compare six models:

a) *mBERT*: [6] pre-trained on Wikipedia in the 104 most frequent languages with two objectives: (1) Masked language modeling (MLM); predicting randomly masked out words and (2) next sentence prediction (NSP) determining whether two

sentences are consecutive only using the representation of the [CLS] token (Fig. 1). It was not trained on any parallel text.

b) *XLM-R*: [16] pre-trained on CommonCrawl (which contains Wikipedia) in 100 languages. It is based on RoBERTa [9], so it only pre-trains with the MLM objective. It was not trained on any parallel text.

c) *XLM-15 (MLM+TLM)*: [34] pre-trained on Wikipedia in the 15 languages from the XNLI dataset [35]. It is pre-trained on MLM objective but also on parallel text (drawn from XNLI) using the Translated Language Modeling (TLM) objective, which is an MLM objective applied to parallel text, to allow the model to attend to words from the other language to predict masked out words. Additionally to its training on parallel data, its input embedding is added to a language embedding indicating the language of the sentence, whereas models like mBERT and XLM-R have no input information about the language.

d) *XLM-100*: [34] pre-trained on Wikipedia in 100 languages with MLM only.

e) *AWESOME*: [32] which is mBERT fine-tuned on a variety of self-supervised objectives and supervised objectives on a parallel corpus to improve word-level alignment for extracting pairs of translated words in parallel sentences: MLM, TLM but also objectives on the consistency of the produced alignment.

f) *mBART*: [36] contrary to all previously mentioned models which are Transformer encoders, mBART follows an encoder-decoder architecture. It was pre-trained on filling missing spans of texts for 50 languages. We consider only the representations built by the encoder part of the model as we empirically observed that the decoder gives worse multilingual alignment than the encoder.

To evaluate the multilingual alignment of those models, we rely on the WMT19 dataset [37] for parallel sentences. And for the bilingual dictionaries, we use MUSE [23]. Monolingual FastText embeddings [26] aligned with RCSLS [24] are used as a baseline.

For all experiments, the number of sampled pairs is $N = 5000$ as in [2]. The number of neighbors for the CSLS criterion (cf. Eq. 2) is $k = 10$ as in [24]. To avoid favoring contextualized models over FastText aligned embedding we chose to sample distinct pairs of words. We empirically verified for all the layers of three models (mBERT, XLM-R, AWESOME) on three language pairs and for 10 different sampling of pairs of words that it gives equivalent results: we observed a strong correlation with a 0.86 Spearman rank correlation (p-value < 0.01). To obtain the 95% confidence intervals on all figures and the empirical standard deviation for all tables, we perform 10 runs of each experiment.

V. RESULTS

In this section, we first investigate the fact that different results were obtained regarding sentence-level alignment. Then, we demonstrate that our method provides better pair of words than FastAlign thanks to a lesser number of carefully selected pairs. Finally, we perform the word-level evaluation

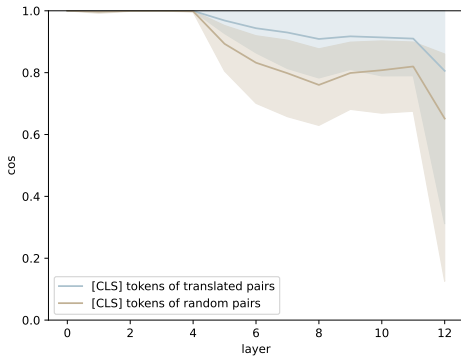


Fig. 4. Evolution of the similarity between CLS tokens across layers of mBERT for translated and random pairs (ru-en), with 95% confidence interval.

allowed by our method for weak alignment first and then for strong alignment, showing that multilingual Transformer-based models bring better alignment than multilingual word embeddings according to our metrics.

A. The chosen sentence representation influences the results

Before reporting our results on word-level alignment, we investigate the contradiction between [2] and [3] on sentence-level alignment for the mBERT model.

On the one hand, [2] performed a nearest-neighbor search similar to ours on sentence representations. Each sentence was represented by the average of the embeddings of its tokens and those vectors were centered for each language. High retrieval accuracy is observed with this method for typologically similar languages.

On the other hand, [3] reported a Canonical Correlation Analysis (CCA) across layers of the representation in various contexts of the initial token [CLS] which is expected to encode the meaning of the sentence. This method shows that those representations are dissimilar, and the dissimilarity grows stronger towards the deeper layers.

On Fig. 4 we observe the same decrease across layers, as [3], for the similarity between [CLS] tokens of translated sentences with the cosine similarity instead of the CCA. But the similarity decreases even more for an equal number of random pairs of sentences drawn from the same dataset.

The decrease of the similarity between translated pairs can be deduced from the fact that it reaches exactly 1 at the 0th layer, corresponding to non-contextualized embeddings which will have the same value if the token is the same, here "[CLS]". The similarity can do nothing but decrease when information from the context is injected into the contextualized representation of the [CLS] token.

However, to fairly compare representations based on the CLS token and averaged representations of the sub-words, they can both be used in the same nearest-neighbor task as proposed in our method for word representations. We apply Equation 1 to sentence representations. The sentence representations are not centered as in [2] as we want to evaluate directly the quality of alignment, and not to artificially improve it.

TABLE I
NN-SEARCH FOR SENTENCE REPRESENTATIONS

method and layer	de-en	ru-en	zh-en
FastText avg	53.3 (1.1)	26.1 (1.1)	1.4 (0.2)
CLS first	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
avg first	56.1 (2.1)	17.8 (0.7)	41.7 (1.1)
CLS best	77.9 (0.4)	59.6 (0.5)	51.1 (0.5)
avg best	90.1 (0.2)	82.1 (0.4)	88.1 (0.6)
CLS last	60.9 (0.8)	40.5 (0.5)	21.0 (0.8)
avg last	87.3 (0.4)	75.5 (0.5)	79.4 (0.3)

TABLE II
PRECISION OF THE EXTRACTED PAIRS

method	en-de	en-fr	ro-en
ours	90.1	95.2	94.5
FastAlign	71.3	80.0	71.8

Results reported in Table I show that whatever the chosen layer is, averaged representations of the sentences provide better multilingual alignment than CLS representations. And for deeper layers, both give better alignment than aligned word embeddings averaged over the sentence. It must also be noted that alignment of the averaged mBERT representation suffers less from the typological distance between languages than the CLS representation or aligned FastText.

Sentence representations can give different results according to the chosen method, but it seems that whatever it is, sentence representations produced by mBERT are relatively aligned across languages. However, the CLS representation seems to be less relevant. Furthermore, the CLS token does not exist in all Transformer-based models, hence we recommend relying on averaged representations instead.

B. Using bilingual dictionary over alignment tools

Fair multilingual alignment of sentence representations does not guarantee a good word-level multilingual alignment. In Section III, we argued that the proposed method is a way to extract translated words which is less prone to errors. Table II shows the proportion of accurate pairs extracted by our method and FastAlign [4]. It demonstrates that our method extracts proportionally more accurate pairs than FastAlign, although it provides fewer pairs in quantity to be fair. For 10 000 sentences from WMT19 for the English-German pair: our method extracts 50 590 word pairs and 190 665 for FastAlign.

[5] used FastAlign to compare the similarity of translated pairs of words and random ones with mBERT. When measuring that similarity on the last layer and plotting the sampled distribution, they observe that the pairs obtained with FastAlign give a very broad distribution that overlaps a lot with random pairs, which leads them to conclude that word-level representations built by mBERT are not well aligned across languages and it motivates them to propose a method to re-align representation after pre-training.

But because they use FastAlign, they are considering many unrelated pairs as translations. Fig. 5 shows those distributions for the eighth layer of mBERT. With its right and wrong extracted pairs, the distribution of FastAlign pairs (in blue)

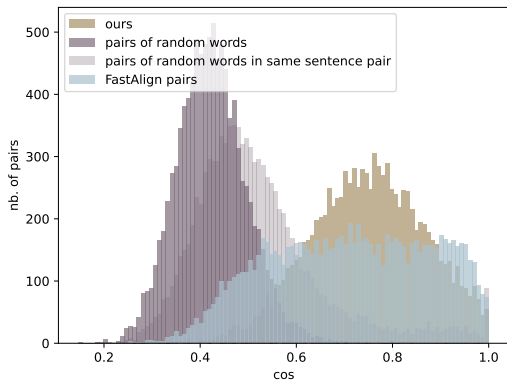


Fig. 5. Distribution of similarity between various types of pairs (en-de).

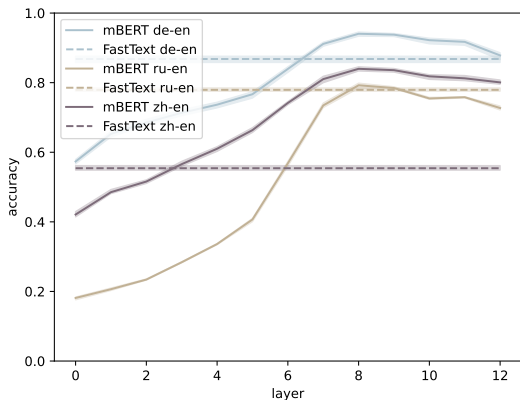


Fig. 6. Evolution of S_{weak} across layers for mBERT compared with FastText aligned.

overlaps more with the distribution of random pairs (in purple) than the distribution of pairs extracted with our methods (in yellow). The distribution of random pairs from the same sentence pair (in light purple) is very close to the distribution of random pairs in the whole dataset (in dark purple), which justifies that any error in the extraction of a pair might increase the overlap between extracted pairs and random pairs.

This confirms that FastAlign generates too many mistakes to make an accurate evaluation of the multilingual alignment produced by a model.

C. Multilingual Encoders produce good word-level alignment

Having advocated for our method, we perform the nearest-neighbor search described in Equation 1 for evaluating the multilingual alignment. Results for mBERT and five language pairs are shown in Fig. 6. We observe that for a few deep layers, mBERT produces representations that are better aligned than multilingual word embeddings.

Lower layers give worse alignment. This implies that multilingual patterns are high-level features. It also goes against the hypothesis made by several papers [2], [3], [19], that shared vocabulary is what allows models like mBERT to align representations without having been exposed to parallel texts.

TABLE III
NN-SEARCH RESULTS FOR MORE MODELS

layer	model	de-en	ru-en	zh-en
-	FastText	86.8 (0.67)	77.9 (0.39)	55.4 (0.43)
best	mBERT	94.1 (0.47)	79.2 (0.64)	84.0 (0.40)
	XLNet	83.5 (0.52)	67.4 (0.38)	27.4 (0.39)
	XLNet Base	87.7 (0.41)	68.7 (0.24)	63.6 (0.50)
	XLNet Large	88.9 (0.30)	76.8 (0.15)	72.3 (0.36)
	XLNet ^{a,b}	68.5 (0.31)	28.4 (0.30)	24.5 (0.56)
	AWESOME ^a	93.4 (0.53)	76.1 (0.60)	82.7 (0.38)
last	mBERT ^{b,c}	92.1 (0.52)	81.2 (0.79)	74.7 (0.34)
	mBERT	87.9 (0.74)	72.7 (0.50)	80.1 (0.43)
	XLNet	82.4 (0.60)	64.6 (0.43)	25.7 (0.46)
	XLNet Base	77.2 (0.81)	49.3 (0.68)	51.0 (0.35)
	XLNet Large	77.4 (0.49)	53.5 (0.65)	54.4 (0.43)
	XLNet ^{a,b}	28.5 (0.60)	4.6 (0.23)	11.2 (0.39)
last	AWESOME ^a	86.5 (0.47)	67.4 (0.21)	79.4 (0.44)
	mBERT ^{b,c}	92.1 (0.52)	81.2 (0.79)	74.7 (0.34)

^a uses parallel data in pre-training ^b encodes the language in input ^c encoder-decoder model (we only evaluate the encoder)

The very last layers also give worse results than layers 8 to 10. [19] have shown that mBERT representations hold language-specific information at each layer. These language-specific components might take more importance in the last layers as the pre-training objective is to predict masked words, a language-specific task. Indeed, the model must learn not to replace a masked-out word with its translation.

This type of curve for S_{weak} is observed for all the multilingual models we evaluated. Table III show retrieval scores for the first, best and last layer of models described in Section IV.

The best layer of mBERT gives the best results with respect to all other models. It might be due to the fact that the next sentence prediction task helps, an objective on which it is the only pre-trained model. However, it could also be explained by the fact that mBERT is trained solely on Wikipedia whereas models like XLNet are trained on the CommonCrawl corpus which might contain texts that are less comparable across languages. It is also to be noted that the TLM objective on parallel texts proposed by the XLNet model seems to make the alignment worse but it is also used in AWESOME.

As the different evaluated models have many differences and obtain somewhat similar results, one cannot isolate a single parameter that makes the multilingual alignment better or worse. Nevertheless, it seems that most of those multilingual models build multilingual representations that are competitive with word embeddings that have been explicitly aligned. Further research is needed for identifying the factors that make the quality of such alignment.

D. Multilingual Encoders produce 'strong' alignment

Finally, the same models are evaluated for the strong alignment retrieval criterion S_{strong} defined in Equation 3. Results for mBERT are reported on Fig. 7 and results for all models are reported on Table IV.

The multilingual alignment of most of the models seems to be robust. Retrieval accuracy is significantly greater for those multilingual models than for multilingual word embeddings.

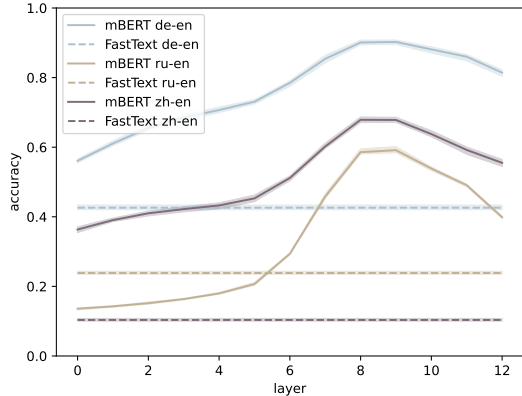


Fig. 7. mBERT and FastText aligned representation evaluated with S_{strong}

TABLE IV
NN-SEARCH RESULTS FOR STRONG ALIGNMENT

layer	model	de-en	ru-en	zh-en
-	FastText	42.6 (0.47)	23.8 (0.33)	10.4 (0.23)
best	mBERT	90.3 (0.47)	59.1 (0.68)	67.9 (0.49)
	XLM-100	81.2 (0.49)	57.9 (0.58)	22.5 (0.66)
	XLM-R Base	82.9 (0.46)	53.5 (0.69)	49.7 (0.45)
	XLM-R Large	87.6 (0.40)	70.4 (0.61)	65.1 (0.47)
	XLM-15 ^{a,b}	62.3 (0.63)	16.7 (0.28)	21.6 (0.58)
	AWESOME ^a	91.6 (0.82)	64.6 (0.63)	70.9 (0.40)
	mBART ^{b,c}	88.4 (0.45)	68.0 (0.82)	60.0 (0.58)
last	mBERT	81.5 (0.63)	39.9 (0.27)	55.5 (0.65)
	XLM-100	72.9 (0.37)	39.0 (0.58)	18.6 (0.55)
	XLM-R Base	73.6 (0.49)	36.4 (0.65)	38.6 (0.39)
	XLM-R Large	72.2 (0.41)	40.5 (0.50)	42.6 (0.48)
	XLM-15 ^{a,b}	20.2 (0.79)	3.7 (0.20)	8.4 (0.48)
	AWESOME ^a	80.2 (0.77)	40.5 (0.18)	56.9 (0.33)
	mBART ^{b,c}	88.4 (0.45)	68.0 (0.82)	60.0 (0.58)

^a uses parallel data in pre-training ^bencodes the language in input
^c encoder-decoder model (we only evaluate the encoder)

There is yet again no way to identify what makes one Transformer-base model perform better than another. Nevertheless, we have demonstrated that there is a word-level strong alignment in most multilingual Transformer-based language models, even for those like mBERT and XLM-R which have no explicit information about the language in input and haven't been pre-trained on parallel texts.

VI. DISCUSSION

From the previous analysis, several questions remain, such as what types of errors are made and whether this performance is a result of real alignment or the product of varying densities of sentences of diverse domains. To answer these questions, we studied a random sample of incorrect predictions on the WMT19 German-English dataset. As shown in Table V, several error categories are highlighted. The first is due to similar lexical fields between the target and prediction, which shows that semantically similar examples are clustered together. The second error category shows that context (e.g. political discourse) can be more important than word representation, especially for words with low information content, such as

TABLE V
QUALITATIVE ANALYSIS OF ALIGNMENT MISTAKES ON EXAMPLES FROM WMT19 FOR THE ENGLISH-GERMAN PAIR

reference word	target word	model output	error category
angehende	aspiring	genuine	lexical field
Anleihe	bond	payout	lexical field
Kapitel	chapter	paragraph	lexical field
Köpfe	heads	skeletons	lexical field
Seen	lakes	fjords	lexical field
gestarkt	strengthened	guaranteed	context ("European Union")
denke	think	deny	context ("Canada", "European")
Bligh	Bligh	Blimp	proper noun
Rohrzucker	cane	soldering	
Inspektion	inspection	persuasion	
standard	default	exclusive	

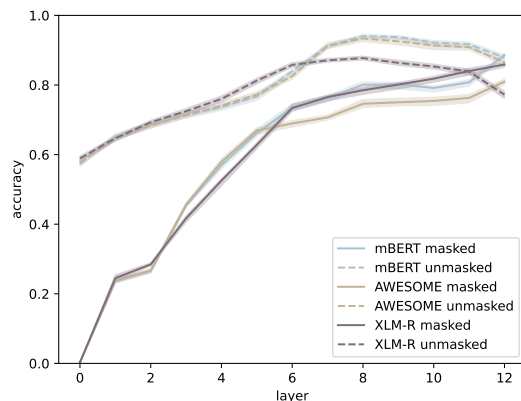


Fig. 8. Top-1 accuracy on WMT19 German-English with masked targets

"think". Third, proper nouns are harder to translate, and they are sometimes associated with other random nouns. Finally, other error patterns are harder to pinpoint and may be due to density problems.

These considerations suggest how important context is when predicting a translation. To show this, we studied the performance of models when replacing the reference and target words with the model mask token, as shown in Figure 8. Although performance deteriorates when compared to performance on the complete sentence, most models perform relatively well without important information (with a decrease of around 15 points). This suggests that context is indeed very useful when predicting translations. Since the tested models are trained with an MLM task, each layer refines the representation of the mask token, which might explain why performance is increasing with the depth of representation, contrary to the previous behaviors. Furthermore, the satisfactory performance on this task suggests that the extracted representation would be well suited for a cross-lingual retrieval task.

VII. CONCLUSION

Our results show that Transformer-based models like mBERT produce contextualized representations of words that

are well aligned across languages, particularly in the deeper layers, despite those models having only been trained on monolingual objectives. By comparing results for weak and strong alignment, it was also shown that multilingual Transformers perform far better on a more challenging evaluation than cross-lingual embeddings which were built with explicit cross-lingual training signal, like FastText embeddings aligned with RCSLS but also deeper models trained with TLM objective such as XLM-15. Finally, our experiments show that averaged representations are better aligned than representations based on CLS token.

For future works, we plan to evaluate the extracted representation on retrieval tasks, as our results show that the studied models seem to be particularly well suited for such a task. Furthermore, these results are promising for multilingual Named Entity Linking, which is closely related to a retrieval task, and could take advantage of data in multiple languages.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, vol. 30. Curran Associates, Inc., 2017.
- [2] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proceedings of ACL*. ACL, Jul. 2019.
- [3] J. Singh, B. McCann, R. Socher, and C. Xiong, "Bert is not an interlingua and the bias of tokenization," in *EMNLP*, 2019.
- [4] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of ibm model 2," in *NAACL*, 2013.
- [5] W. Zhao, S. Eger, J. Bjerva, and I. Augenstein, "Inducing language-agnostic multilingual representations," *CoRR*, vol. abs/2008.09112, 2020.
- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL 2019*, Jun. 2019.
- [7] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," 2020.
- [8] E. M. Bender, "On achieving and evaluating language-independence in nlp," *Linguistic Issues in Language Technology*, Oct. 2011.
- [9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2020.
- [10] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: Pretrained language model for scientific text," in *EMNLP*, 2019.
- [11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [12] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of ACL 2020*. Online: ACL, Jul. 2020.
- [13] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, "FlauBERT: Unsupervised language model pre-training for French," in *Proceedings of LREC*. ELRA, May 2020.
- [14] S. Rönqvist, J. Kanerva, T. Salakoski, and F. Ginter, "Is multilingual BERT fluent in language generation?" in *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*. Turku, Finland: Linköping University Electronic Press, Sep. 2019.
- [15] S. Ralethe, "Adaptation of deep bidirectional transformers for Afrikaans language," in *Proceedings of LREC*. ELRA, May 2020.
- [16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of ACL*. Online: ACL, Jul. 2020.
- [17] N. Chomsky, *Language and Mind*. Harcourt Brace, 1968.
- [18] J. H. Greenberg, *Language universals*. Mouton The Hague, 1966.
- [19] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," in *Proceedings of EMNLP-IJCNLP*. Hong Kong, China: ACL, Nov. 2019.
- [20] S. Doddapaneni, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "A primer on pretrained multilingual language models," *CoRR*, vol. abs/2107.00676, 2021.
- [21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.
- [22] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *CoRR*, vol. abs/1309.4168, 2013.
- [23] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.
- [24] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," in *Proceedings of EMNLP*. ACL, Oct.-Nov. 2018.
- [25] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *Proceedings of ACL*. ACL, Jul. 2018.
- [26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, 2016. [Online]. Available: <http://arxiv.org/abs/1607.04606>
- [27] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," *CoRR*, vol. abs/1804.07755, 2018.
- [28] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *CoRR*, vol. abs/1812.10464, 2018.
- [29] J. Eisenschlos, S. Ruder, P. Czaplá, M. Kadras, S. Gugger, and J. Howard, "MultiFiT: Efficient multi-lingual language model fine-tuning," in *Proceedings of EMNLP-IJCNLP*. ACL, Nov. 2019.
- [30] T. Limisiewicz, D. Mareček, and R. Rosa, "Universal Dependencies According to BERT: Both More Specific and More General," in *Findings of EMNLP 2020*. Online: ACL, Nov. 2020.
- [31] R. Choenni and E. Shutova, "What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties," *CoRR*, vol. abs/2009.12862, 2020.
- [32] Z. Dou and G. Neubig, "Word alignment by fine-tuning embeddings on parallel corpora," *CoRR*, vol. abs/2101.08231, 2021.
- [33] U. Roy, N. Constant, R. Al-Rfou, A. Barua, A. Phillips, and Y. Yang, "LAREQA: Language-agnostic answer retrieval from a multilingual pool," in *Proceedings of EMNLP*. ACL, Nov. 2020.
- [34] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *CoRR*, vol. abs/1901.07291, 2019.
- [35] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: evaluating cross-lingual sentence representations," *CoRR*, vol. abs/1809.05053, 2018.
- [36] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *CoRR*, vol. abs/2001.08210, 2020.
- [37] W. Foundation. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news. [Online]. Available: <http://www.statmt.org/wmt19/translation-task.html>