Language Representation in Multilingual BERT and its applications to improve Cross-lingual Generalization

Chi-Liang Liu* **Tsung-Yuan Hsu*** **Yung-Sung Chuang** **Chung-Yi Li** **Hung-Yi Lee**
College of Electrical Engineering and Computer Science, National Taiwan University
{liangtaiwan1230, sivia89024, tlkagkb93901106}@gmail.com
{b05901033, r07942080}@ntu.edu.tw

Abstract

A token embedding in multilingual BERT (m-BERT) contains both language and semantic information. We find that representation of a language can be obtained by simply averaging the embeddings of the tokens of the language. With the language representation, we can control the output languages of multilingual BERT by manipulating the token embeddings and achieve *unsupervised token translation*. We further propose a computationally cheap but effective approach to improve the cross-lingual ability of m-BERT based on the observation.

1 Introduction

Multilingual BERT (m-BERT) (Devlin et al., 2019) has already demonstrated its extraordinary strength in cross-lingual transfer on a variety of tasks (Conneau et al., 2018; Wu and Dredze, 2019; Hsu et al., 2019; Pires et al., 2019), and this has been credited to the cross-lingual alignment of its internal representations, which is a phenomenon that semantically-similar or functionally-similar words from different languages are represented with similar embeddings.

Although m-BERT does exceptionally well in aligning cross-lingual representation, we believe the language information is still in token embeddings. One evidence is that m-BERT learns to reconstruct the same sentences as input, so when masking one token from an English sentence, the probability of decoding a Chinese token in that position is very low. Hence, m-BERT reserves some implicit language information in the embedding space that can be disentangled from semantic information. To verify the assumption, we show that if an English sentence is given to m-BERT as input, and then its embeddings are shifted in a specific direction in the embedding space, then m-BERT would output a sentence in another language semantically close to the input sentence.

After showing the existence of language information in the embeddings of m-BERT, we eliminate these language-specific variations in the embeddings, and demonstrate that its a practical way to boost zero-shot cross-lingual transferability of m-BERT on downstream tasks.

In literature, there are studies trying to improve cross-lingual alignment in pre-trained m-BERT. For example, Cao et al. (2020) proposed to fine-tune m-BERT on a small parallel data and Libovický et al. (2020) proposed to unsupervisedly zero-centering the embeddings language by language to achieve language neutrality and showed progress on retrieval tasks. Our work is concurrent to Libovický et al. (2020)'s and has similarities in approaches. But the approach here is more effective.

The contributions of this work can be summarized as the following:

- The language information in m-BERT can be represented by the average of all token embeddings of the specific language. This is verified by *unsupervised token translation*.
- The cross-lingual transferability of m-BERT in downstream tasks can be improved by manipulating token embeddings.
- We compared our method to other cross-lingual embedding alignment methods.

2 Language Representation

We assumed that we have n languages denoted by $\{L_1, L_2, \ldots, L_n\}$ and their corresponding corpora.

*Equal Contribution

[0]This work was supported by Delta Electronics, Inc..

Context-Dependent Representation Given an input sequence x and token index i, we denote the hidden representation in layer l by $\boldsymbol{h}^l_{x_i}$.

Mean of Language Given a language L and its corresponding corpora C contain a set of sentences x, we denote the mean of language of layer l by

$$\boldsymbol{R}^l_L = \mathop{\mathbb{E}}_{x_i \in C} \left[\boldsymbol{h}^l_{x_i} \right],$$

which represents the mean of all the token embeddings in the corpora. We assume that mean of language contains language-specific information but no semantic information.

Although the assumption about language representation here seems naive, in the experiments, we show that $\boldsymbol{R}^l_L$ represents the language information in token embeddings well. For each language L, there is a language-specific representation $\boldsymbol{R}^l_L$ for each layer l. Because we do not know $\boldsymbol{R}^l_L$ from which layer l represents language L the best, l is a hyperparameter in the following algorithms.

Zero-mean We can simply subtract each token embedding $\boldsymbol{h}^l_{x_i}$ by mean of language $\boldsymbol{R}^l_{L_k}$ to eliminate language-specific information. The subtraction moves the token embedding to a language-agnostic joint space. The language-agnostic hidden representation $\hat{\boldsymbol{h}}^l_{x_i}$ can be written as

$$\hat{\boldsymbol{h}}^l_{x_i} = \boldsymbol{h}^l_{x_i} - \boldsymbol{R}^l_{L_k}, \tag{1}$$

where $\boldsymbol{h}^l_{x_i}$ is extracted from the token x_i in L_k.

Mean Difference Shift (MDS) Besides eliminating language information, we can move the embedding in the space of L_1 to the space of L_2, or *unsupervised token translation*. That is, given the embedding of *me* in English, we can modify the embedding to make it be interpreted by m-BERT as the embedding of 我(*me* in Chinese).

We fed an L_1 sentence into m-BERT, extracted the embedding of each token at layer l. Then we subtract the embedding by $\boldsymbol{R}^l_{L_1}$ as in (1) to remove the information of L_1, and then we add $\boldsymbol{R}^l_{L_2}$ to shift the embedding to the space of L_2. Formally, we modified token embedding $\boldsymbol{h}^l_{x_i}$ in $L1$ into embedding $\tilde{\boldsymbol{h}}^l_{x_i}$ in $L2$ as below:

$$\tilde{\boldsymbol{h}}^l_{x_i} = \boldsymbol{h}^l_{x_i} + \boldsymbol{R}^l_{L_2} - \boldsymbol{R}^l_{L_1}. \tag{2}$$

3 Unsupervised Token Translation

In this section, we showed that the implicit language-specific information in the embedding space could be disentangled from semantic embeddings. We used MDS to make m-BERT take a sentence in L_1 as input, and *translate* it into a sentence in L_2.

3.1 Setup

The formulation of MDS was modified slightly from (2): $\tilde{\boldsymbol{h}}^l_{x_i} = \boldsymbol{h}^l_{x_i} + \alpha \left(\boldsymbol{R}^l_{L_2} - \boldsymbol{R}^l_{L_1} \right)$, where α was a hyperparameter, and we will see its influence later in the experimental results. Given the input, the token embeddings were modified at a specific layer l. The $l+1$-th layer took the modified embeddings as input, and the final layer generated a sequence of tokens. The sentences in this experiment were from XNLI test-set, which contains 15 languages, including low resource languages such as Swahili and Urdu.

3.2 Evaluation metrics

We use two different metrics to analyze the results of unsupervised token translation quantitatively.

BLEU-1 score This metric measures the translation quality without considering the fluency of the converted sequence.

Convert Rate Besides translation quality, we also calculated convert rate, the percentage of tokens converted from source language to target language, which was defined as follow:

$$\text{convert rate} = \frac{\text{\# of } y \in (V_t - V_s)}{\text{\# of } y - \text{\# of } y \in V_s \cap V_t},$$

where y is the output tokens of the model, V_s, V_t is the token set of the source and target language. Shared tokens in both vocabularies were not taken into account therefore excluded from the numerator and denominator term.

3.3 Results

Surprisingly, we were able to get predicted tokens in language L_2 from L_1 input by applying MDS, and many of the predicted tokens were the token-level translation of the input tokens in L_1, even on the low resource languages. The sample outputs were shown in Appendix A.

Table 1 shows the quantitative results. First, although the translation results were not comparable to the existing unsupervised translation methods (Kim et al., 2018), it showed strong evidence that we could manipulate the language-specific information in the token embedding space by MDS,

Table 1: Unsupervised Token Translation quantitative results using the 10-th layer of BERT.

	en→de	en→fr	en→ur	en→sw	en→zh	en→el	de→en	fr→en	ur→en	sw→en	zh→en	el→en
BLEU-1 (α=1)	7.53	8.53	5.56	7.96	15.25	7.88	7.34	9.08	5.52	6.34	4.37	6.54
BLEU-1 (α=2)	8.03	10.24	6.31	7.23	21.51	14.91	7.48	8.52	6.23	7.48	5.38	6.65
BLEU-1 (α=3)	12.35	10.65	5.35	7.16	15.95	19.13	6.29	12.27	5.74	6.45	6.17	4.73
convert rate (α=1)	40.2	41.7	61.1	15.3	47.8	62.1	45.2	49.6	29.9	14.7	23.9	30.2
convert rate (α=2)	74.8	75.7	99.4	97.4	90.0	99.1	67.3	601	83.0	65.6	60.8	97.9
convert rate (α=3)	95.2	96.3	99.8	100	99.5	100	79.5	73.1	96.6	93.6	91.4	99.7

Table 2: Sentence Retrieval Result on Tatoeba using the 8-th layer of BERT.

Method	de	es	ar	el	fr	hi
Original	75.4	64.1	24.5	29.8	64.3	**34.8**
Zero-mean	73.5	61.8	23.5	29.4	63.7	29.9
MDS	**76.8**	**67.5**	**29.1**	**30.6**	**67.0**	31.4
	ru	vi	th	tr	zh	
Original	**63.6**	**61.0**	13.7	32.9	68.6	
Zero-mean	59.6	51.2	13.7	32.8	64.1	
MDS	59.4	51.5	**17.5**	**36.8**	**69.2**	

Table 3: Sentence Retrieval Result on BUCC2018 dev-set and test-set using the 8-th layer of BERT.

	Method	de	fr	ru	zh
dev	Original	75.62	72.07	68.59	66.04
	Zero-mean	71.10	70.51	65.92	59.91
	MDS	**76.91**	**73.45**	**71.60**	**66.91**
test	Original	63.22	62.47	11.65	50.47
	Zero-mean	59.59	59.25	10.40	45.42
	MDS	**65.76**	**63.95**	**12.36**	**52.45**

and force m-BERT to switch from one language to another. Second, we observed that as α became larger, the model converted more tokens to target language L_2 and never decoded tokens not belonging to L_1 and L_2. When given a negative α, the model always decoded tokens belonging to L_1. The above observation showed that in the embedding space, the direction related to language is unique. We showed a further analysis of α in Appendix A.

4 Cross-lingual Sentence Retrieval

Extracting parallel sentences from a comparable corpus between two languages is one of the standard methods of evaluating cross-lingual embeddings (Hu et al., 2020; Zweigenbaum et al., 2017; Artetxe and Schwenk, 2018). In this section, we demonstrated that embeddings shifted by MDS are better cross-lingually aligned through evaluations on a sentence-level retrieval task.

4.1 Task

We evaluated the effect of MDS and Zero-mean on two sentence retrieval tasks: BUCC2018 and Tatoeba. We used the mean vector of all token embeddings in a sentence as the sentence embedding and cosine similarity as distance metric. Token embeddings were extracted from a specific layer of BERT encoder, and MDS or Zero-mean shifts pre-computed on the whole dataset were applied directly to the extracted embeddings.

4.2 MDS vs Zero-Mean

Although applying MDS or applying Zero-mean in sentence retrieval task seem very similar at first glance, they have subtle differences. Assume we have sentence embeddings $\boldsymbol{v_1} \in L_1$ and $\boldsymbol{v_2} \in L_2$, and the two sentences in different languages have the same semantic meaning. Assume there exist *real* language representations $\boldsymbol{R}^*_{L_1}$ and $\boldsymbol{R}^*_{L_2}$ that can perfectly eliminate language from embedding[1], leading to $\boldsymbol{v_1} - \boldsymbol{R}^*_{L_1} = \boldsymbol{v_2} - \boldsymbol{R}^*_{L_2}$. The language representations $\boldsymbol{R}_{L_1}$ and $\boldsymbol{R}_{L_2}$ obtained by average are the approximation of the real ones[2], and δ_1 and δ_2 are difference between real and approximate language representations.

$$\begin{aligned} &\boldsymbol{v_1} - \boldsymbol{R}_{L_1} \neq \boldsymbol{v_2} - \boldsymbol{R}_{L_2} \\ \rightarrow &\boldsymbol{v_1} - \boldsymbol{R}_{L_1} - \boldsymbol{\delta_1} = \boldsymbol{v_2} - \boldsymbol{R}_{L_2} - \boldsymbol{\delta_2} \end{aligned}$$

Then the cosine similarities of $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ after Zero-mean or MDS are shown as below,

$$\begin{aligned} \cos_{\text{MDS}}(\boldsymbol{v_1}, \boldsymbol{v_2}) &= \frac{(\boldsymbol{v_1} - \boldsymbol{R_{L_1}} + \boldsymbol{R_{L_2}}) \cdot \boldsymbol{v_2}}{|\boldsymbol{v_1} - \boldsymbol{R_{L_1}} + \boldsymbol{R_{L_2}}|\,|\boldsymbol{v_2}|} \\ &= \frac{(\boldsymbol{v_2} + \boldsymbol{\delta}) \cdot \boldsymbol{v_2}}{|\boldsymbol{v_2} + \boldsymbol{\delta}|\,|\boldsymbol{v_2}|},\ \text{where } \boldsymbol{\delta} = \boldsymbol{\delta_1} - \boldsymbol{\delta_2} \\ \cos_{\text{Zero-mean}}(\boldsymbol{v_1}, \boldsymbol{v_2}) &= \frac{(\boldsymbol{v_1} - \boldsymbol{R_{L_1}}) \cdot (\boldsymbol{v_2} - \boldsymbol{R_{L_2}})}{|\boldsymbol{v_1} - \boldsymbol{R_{L_1}}|\,|\boldsymbol{v_2} - \boldsymbol{R_{L_2}}|} \\ &= \frac{(\boldsymbol{v_2} - \boldsymbol{R^*_{L_2}} + \boldsymbol{\delta_1})\,(\boldsymbol{v_2} - \boldsymbol{R^*_{L_2}} + \boldsymbol{\delta_2})}{|\boldsymbol{v_2} - \boldsymbol{R^*_{L_2}} + \boldsymbol{\delta_1}|\,|\boldsymbol{v_2} - \boldsymbol{R^*_{L_2}} + \boldsymbol{\delta_2}|} \end{aligned}$$

The above result shows that Zero-mean method is more sensitive to the approximate error when $|\boldsymbol{v_2}| > |\boldsymbol{v_2} - \boldsymbol{R_{L_2}}| > \max(|\boldsymbol{\delta_1}, |\boldsymbol{\delta_2}|, |\boldsymbol{\delta}||)$.[3] The

[1] Unknown to us

[2] The superscript ^{l} is ignored here for simplicity

[3] This is very possible. Because $\boldsymbol{v_2}$ is in L_2, it may has the same direction as $\boldsymbol{R}_{L_2}$.

differences between the two methods were further verified in the experiments.

4.3 Result

Results of sentence retrieval were shown as table 3 and table 2. On BUCC2018 dev-set and test-set, embeddings shifted by MDS consistently achieved higher accuracies on all of the languages, and Zero-meaned embeddings were worse than doing nothing. On Tatoeba test-set, MDS embeddings were also the best in most of the languages except for Hindi, Russian and Vietnamese.

5 Cross-Lingual Transfer

5.1 Setup

In zero-shot cross-lingual transfer learning, m-BERT was fine-tuned on the source language, which was English in the following experiments, and tested on languages never seen during fine-tuning. For each language, we had around 5M tokens from Wikipedia documents to compute the language representations.

Zero-mean During fine-tuning, we applied Zero-mean on the token embeddings of the source language and forwarding the modified embeddings to the remaining layers during training. While testing, we fed the fine-tuned model with target language data and applied Zero-mean to the embeddings at layer l as well. The means of language vectors were extracted from the Wikipedia data from the **pre-trained** model.

MDS In this approach, we did not need to modify embeddings when training. While testing, we applied MDS to the embeddings at layer l. The mean difference vectors were extracted from Wikipedia data by **fine-tuned** model.

5.2 Tasks

We did the experiments on two tasks, Part-of-Speech (POS) Tagging and Dependency Parsing, to show that our methods could improve cross-lingual zero-shot learning performance.

Part-of-Speech Tagging For POS tagging, we used the Universal Dependencies v2.5 (Nivre et al., 2020) treebanks for 90 languages. Each word was assigned one of 17 universal POS tags. The model was trained on English and tested on 13 other languages.

Dependency Parsing For dependency parsing, we used languages with nonzero amount of training data in CoNLL 2018 Shared Task (Zeman et al., 2018). We evaluated the zero-shot performance on 6 low-resource languages that do not have an official development set. Note that most of the testing languages in dependency parsing had not been included in the pretrained languages of m-BERT[4] except for Kazakh and Irish.

5.3 Results

Table 4 and Table 5 compared the results of baselines and our methods on POS tagging and dependency parsing, respectively. For the POS tagging task, Zero-mean and MDS both improved the performances on the testing sets across languages with only a few exceptions. MDS was not helpful in Thai (th), while both approaches did not improve on Vietnamese (vi). For dependency parsing, although most of the testing languages were not in the pretrained languages of m-BERT, zero-shot transfer learning still achieved reasonable performance. We observed that although Zero-mean does not improve the performance in dependency parsing, MDS improved the performance except for Irish, which showed the effectiveness of MDS.

6 Conclusion

In this paper, we examined the existence of language-specific information in m-BERT embeddings and succeeded in the task of unsupervised token translation by manipulating language-specific information. The proposed methods are further shown to be effective in improving cross-lingual embedding alignment and cross-lingual transfer learning. We will further explore the proposed approach on more downstream tasks.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating

[4] https://github.com/google-research/bert/blob/master/multilingual.md

Table 4: POS Tagging Result.

Method	ar	bg	de	el	es	fr	hi	ru	th	tr	ur	vi	zh	Average
Original	53.8	85.4	86.2	81.1	86.1	42.9	66.8	85.5	41.7	68.6	56.3	**53.8**	61.8	66.9
Zero-mean	54.2	**86.4**	86.5	81.5	**86.8**	**43.9**	**68.9**	86.4	**44.2**	69.4	**57.1**	52.4	**63.0**	**67.8**
MDS	**54.3**	86.1	**86.6**	**81.8**	86.6	43.7	68.1	**86.5**	41.6	**69.7**	56.6	53.4	62.5	67.5

Table 5: Dependency Parsing Result on 6 languages of CoNLL 2018 test set using the last layer of BERT. (Buryat(Bu.), Kurmanji(Ku.), Upper Sorbian(US.), Kazakh(Ka.), Irish(Ir.), North Sami(NS.))

Method	Bu.	Ku.	US.	Ka.	Ir.	NS.	Avg.
Original	16.6	4.9	33.7	39.1	**37.6**	6.7	23.1
Zero-mean	17.1	3.9	31.3	39.6	37.4	6.6	22.7
MDS	**18.2**	**5.9**	**34.2**	**40.0**	37.5	**7.3**	**23.9**

cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. 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 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 862–868.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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.

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.

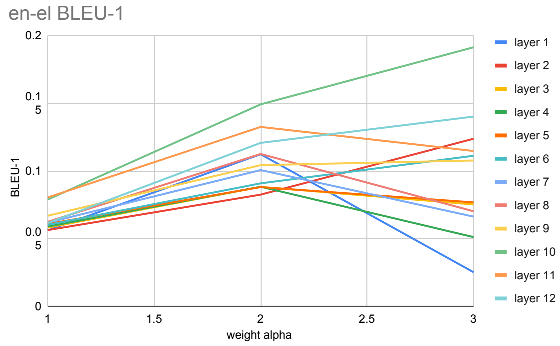
Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

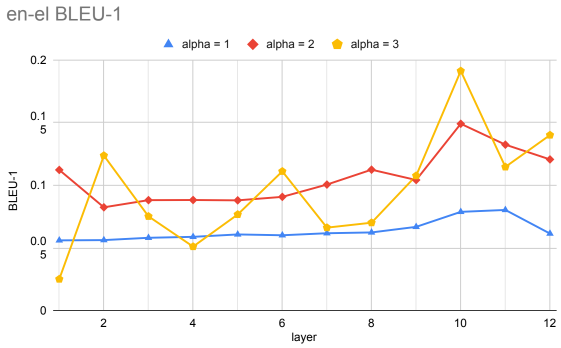
A Analysis of Weight α in Unsupervised Token Translation

Table 6: Size of token set and size of English token set intersection with another language token set.

	en	de	fr	el	zh	ur	sw
$\|V_{\text{lang}}\|$	9140	9212	8552	3189	3866	4085	5609
$\|V_{en} \cap V_{\text{lang}}\|$	9140	3230	3911	1696	1325	1549	2970



(a) x-axis presents different weight α.



(b) x-axis presents different layers.

Figure 1: The direction of change on BLEU-1 of en → el unsupervised token translation when tuning the shifting weight α and selecting different layers.

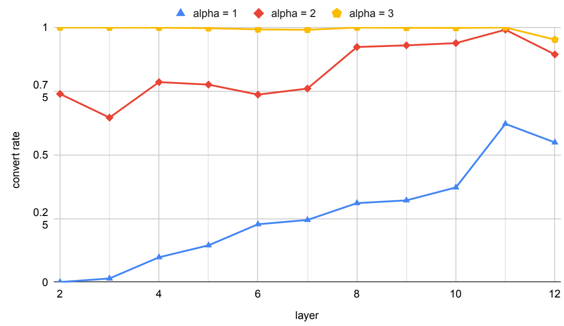


Figure 2: Convert rate on *en → el* data when applying different α on different layers for MDS.

We present an example of *en → el* in an Figure 2 and 1 show that how convert rate and BLEU-1 score changed with different weights α and different layers.

Although, the influences of weight increase on BLEU-1 score were mixed, it is worth mentioning that in the last few layers (10 or 11), the BLEU-1 of most languages obviously rose when α was set to 3.0 (also shown in the best layer row in Table 1). It indicates that the last few layers may be better for disentangling language-specific representations, which is consistent with the observation in the literature that the last few layers contain more language-specific information for predicting masked words (Pires et al., 2019).

Table 7: Unsupervised Token Translation random sample (applied MDS on layer 10)

Input (en)	The girl that can help me is all the way across town. There is no one who can help me.
Ground Truth (zh)	能帮助我的女孩在小镇的另一边。没有人能帮助我。。
en→zh, $\alpha = 1$	. 孩，can 来我是all the way across 市。。There 是无人人can help 我。
en→zh, $\alpha = 2$	. 孩的的家我是这个人的市。。他是他人人的到我。
en→zh, $\alpha = 3$	。，的的的他是的个的的，。：他是他人，的。他。
Ground Truth (fr)	La fille qui peut m'aider est à l'autre bout de la ville. Il n'y a personne qui pourrait m'aider.
en→fr, $\alpha = 1$	. girl qui can help me est all la way across town . . There est no one qui can help me .
en→fr, $\alpha = 2$	. girl qui de help me est all la way dans , . . Il est de seul qui pour aid me .
en→fr, $\alpha = 3$	, , , de , me , all la , , , , , n n n n , , , , ,