First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT

Benjamin Muller[1,2] **Yanai Elazar**[3,4] **Benoît Sagot**[1] **Djamé Seddah**[1]
[1]Inria, Paris, France [2]Sorbonne Université, Paris, France
[3]Computer Science Department, Bar Ilan University
[4]Allen Institute for Artificial Intelligence
{benjamin.muller, benoit.sagot,djame.seddah}@inria.fr
yanaiela@gmail.com

Abstract

Multilingual pretrained language models have demonstrated remarkable zero-shot cross-lingual transfer capabilities. Such transfer emerges by fine-tuning on a task of interest in one language and evaluating on a distinct language, not seen during the fine-tuning. Despite promising results, we still lack a proper understanding of the source of this transfer. Using a novel layer ablation technique and analyses of the model's internal representations, we show that multilingual BERT, a popular multilingual language model, can be viewed as the stacking of two sub-networks: a multilingual encoder followed by a task-specific language-agnostic predictor. While the encoder is crucial for cross-lingual transfer and remains mostly unchanged during fine-tuning, the task predictor has little importance on the transfer and can be reinitialized during fine-tuning. We present extensive experiments with three distinct tasks, seventeen typologically diverse languages and multiple domains to support our hypothesis.

1 Introduction

Zero-shot Cross-Lingual transfer aims at building models for a *target* language by reusing knowledge acquired from a *source* language. Historically, it has been tackled with a two-step *standard cross-lingual pipeline* (Ruder et al., 2019): (1) Building a shared multilingual representation of text, typically by aligning textual representations across languages. This step can be done using feature extraction (Aone and McKee, 1993; Schultz and Waibel, 2001) as with the delexicalized approach (Zeman and Resnik, 2008; Søgaard, 2011) or using word embedding techniques (Mikolov et al., 2013; Smith et al., 2017) by projecting monolingual embeddings onto a shared multilingual embedding space, this step requiring explicit supervision signal in the target language in the form of features or parallel data. (2) Training a task-specific model using supervision on a source language on top of the shared representation.

Recently, the rise of multilingual language models entailed a paradigm shift in this field. Multilingual pretrained language models (Devlin et al., 2019; Conneau and Lample, 2019) have been shown to perform efficient zero-shot cross-lingual transfer for many tasks and languages (Pires et al., 2019; Wu and Dredze, 2019). Such transfer relies on three-steps: (i) pretraining a mask-language model (e.g. Devlin et al. (2019)) on the concatenation of monolingual corpora across multiple languages, (ii) fine-tuning the model on a specific task in the source language, and (iii) using the fine-tuned model on a target language. The success of this approach is remarkable, and in contrast to the standard cross-lingual pipeline, the model sees neither aligned data nor task-specific annotated data in the target language at any training stage.

The source of such a successful transfer is still largely unexplained. Pires et al. (2019) hypothesize that these models learn shared multilingual representations during pretraining. Focusing on syntax, Chi et al. (2020) recently showed that the multilingual version of BERT (mBERT) (Devlin et al., 2019), encodes linguistic properties in shared multilingual sub-spaces. Recently, Gonen et al. (2020) suggest that mBERT learns a language encoding component and an abstract cross-lingual component. In this work, we are interested in understanding the mechanism that leads mBERT to perform zero-shot cross-lingual transfer. More specifically, we ask **what parts of the model and what mechanisms support cross-lingual transfer?**

By combining behavioral and structural analyses (Belinkov et al., 2020), we show that mBERT operates as the stacking of two modules: (1) A multilingual encoder, located in the lower part of the model, critical for cross-lingual transfer, is in charge of

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2214–2231
April 19 - 23, 2021.

aligning multilingual representations; and (2) a task-specific, language-agnostic predictor which has little importance for cross-lingual transfer and is dedicated to performing the downstream task. This mechanism that emerges out-of-the-box, without any explicit supervision, suggests that mBERT behaves like the standard cross-lingual pipeline. Our contributions advance the understanding of multilingual language models and as such have the potential to support the development of better pretraining processes.

2 Analysis Techniques

We study mBERT with a novel behavioral test that disentangles the task fine-tuning influence from the pretraining step (§2.1), and a structural analysis on the intermediate representations (§2.2). Combining the results from these analyses allows us to locate the cross-lingual transfer and gain insights into the mechanisms that enable it.

2.1 Locating Transfer with RANDOM-INIT

In order to disentangle the impact of the pretraining step from the fine-tuning, we propose a new behavioral technique: RANDOM-INIT. First, we randomly initialize a set of parameters (e.g. all the parameters of a given layer) instead of using the parameters learned during the pretraining step. Then, we fine-tune the modified pretrained model and measure the downstream performance.[1]

By replacing a given set of pretrained parameters and fine-tuning the model, *all other factors being equal*, RANDOM-INIT enables us to quantify the contribution of a given set of pretrained parameters on downstream performance and therefore to locate which pretrained parameters contribute to the cross-lingual transfer.

If the cross-language performance is significantly lower than same-language performance, we conclude that these layers are more important to cross-language performance than they are for same-language performance. If the cross-language score does not change, it indicates that cross-language transfer does not rely on these layers.

This technique is reminiscent of the recent *Amnesic Probing* method (Elazar et al., 2020), that removes from the representation a specific feature, e.g. Part-of-Speech, and then measures the outcome on the downstream task. In contrast, RANDOM-INIT allows to study a specific architecture component, instead of specific features.

2.2 Hidden State Similarities across Languages

To strengthen the behavioral evidence brought by RANDOM-INIT, and provide finer analyses that focus on individual layers, we study how the textual representations differ between parallel sentences in different languages. We hypothesize that an efficient fine-tuned model should be able to represent similar sentences in the source and target languages similarly, even-though it was fine-tuned only on the source language.

To measure the similarities of the representation across languages, we use the Central Kernel Alignment metric (CKA), introduced by Kornblith et al. (2019). We follow Conneau et al. (2020) who use the CKA as a similarity metric to compare the representations of monolingual and bilingual pretrained models across languages. In our work, we use the CKA to study the representation difference between source and target languages in pretrained and fine-tuned multilingual models. For every layer, we average all contextualized tokens in a sentence to get a single vector.[2] Then we compute the similarity between target and source representations and compare it across layers in the pretrained and fine-tuned models. We call this metric the *cross-lingual similarity*.

3 Experimental Setup

Tasks, Datasets and Evaluation We consider three tasks covering both syntactic and semantic aspects of language: Part-Of-Speech Tagging (POS), dependency parsing, and Named-Entity Recognition (NER). For POS tagging and parsing we use the Universal Dependency (Nivre et al., 2018) treebanks, and for NER, we use the WikiANN dataset (Pan et al., 2017). We evaluate our systems with the standard metrics per task; word-level accuracy for POS tagging, F1 for NER and labeled attachment score (LAS) for parsing. All the reported scores are computed on the test set of each dataset.

We experiment with English, Russian and Arabic as source languages, and fourteen typologically diverse target languages, including Chinese, Czech, German and Hindi. The complete list can be found in the Appendix A.1.2.

[1] Note that we perform the same optimization procedure for the model with and w/o RANDOM-INIT (optimal learning rate and batch size are chosen with grid-search).

[2] After removing [CLS] and [SEP] special tokens.

SRC-TRG	REF	RANDOM-INIT *of layers* Δ1-2	Δ3-4	Δ5-6	Δ7-8	Δ9-10	Δ11-12
				Parsing			
EN - EN	88.98	-0.96	-0.66	-0.93	-0.55	0.04	-0.09
RU - RU	85.15	-0.82	-1.38	-1.51	-0.86	-0.29	0.18
AR - AR	59.54	-0.78	-2.14	-1.20	-0.67	-0.27	0.08
EN - X	53.23	-15.77	-6.51	-3.39	-1.47	0.29	1.00
RU - X	55.41	-7.69	-3.71	-3.13	-1.70	0.92	0.94
AR - X	27.97	-4.91	-3.17	-1.48	-1.68	-0.36	-0.14
				POS			
EN - EN	96.51	-0.30	-0.25	-0.40	-0.00	0.05	0.02
RU - RU	96.90	-0.52	-0.55	-0.40	-0.07	0.02	-0.03
AR - AR	79.28	-0.35	-0.49	-0.36	-0.19	-0.05	-0.00
EN - X	79.37	-8.94	-2.49	-1.66	-0.88	0.20	-0.14
RU - X	79.25	-10.08	-2.83	-1.65	-2.74	0.01	-0.45
AR - X	64.81	-6.73	-3.50	-1.63	-1.56	-0.73	-1.29
				NER			
EN - EN	83.30	-2.66	-2.14	-1.43	-0.63	-0.23	-0.12
RU - RU	88.20	-2.08	-2.13	-1.52	-0.64	-0.33	-0.13
AR - AR	87.97	-2.37	-2.11	-0.96	-0.39	-0.15	0.21
EN - X	64.17	-8.28	-5.09	-3.07	-0.79	-0.47	-0.13
RU - X	62.13	-15.85	-9.36	-5.50	-2.44	-1.16	-0.06
AR - X	65.59	-16.10	-8.42	-3.73	-1.40	-0.25	0.67

Table 1: Relative Zero shot Cross-Lingual performance of mBERT with RANDOM-INIT (§2.1) on pairs of consecutive layers compared to mBERT without any random-initialization (REF). In SRC-TRG, SRC indicates the source language on which we fine-tune mBERT, and TRG the target language on which we evaluate it. SRC-X is the average across all 17 target language with X $\neq$ SRC. Detailed results per target language are reported in tables 6, 7 and 8 in the Appendix. Coloring is computed based on how mBERT with RANDOM-INIT performs compared to the REF model. $\geq$ REF | $<$ REF | $\leq$ -2 points | $\leq$ -5 points

The results of a model that is fine-tuned and evaluated on the *same* language are referred to as *same-language* and those evaluated on *distinct* languages are referred to as *cross-language*.

Multilingual Model We focus on mBERT (Devlin et al., 2019), a 12-layer model trained on the concatenation of 104 monolingual Wikipedia corpora, including our languages of study.

Fine-Tuning We fine-tune the model for each task following the standard methodology of Devlin et al. (2019). The exact details for reproducing our results can be found in the Appendix. All reported scores are averaged on 5 runs with different seeds.

4 Results

4.1 Disentangling the Pretraining Effect

For each experiment, we measure the impact of randomly-initializing specific layers as the difference between the model performance without any random-initialization (REF) and with random-initialization (RANDOM-INIT). Results for two consecutive layers are shown in Table 1. The rest of the results, which exhibit similar trends, can be found in the Appendix (Table 5).

For all tasks, we observe sharp drops in the cross-language performance at the lower layers of the model but only moderate drops in the same-language performance. For instance, the parsing experiment with English as the source language, results in a performance drop on English of only 0.96 points (EN-EN), when randomly-initializing layers 1 and 2. However, it leads to an average drop of 15.77 points on other languages (EN-X).

Furthermore, we show that applying RANDOM-INIT to the upper layers does not harm same-language and cross-language performances (e.g. when training on parsing for English, the performance slightly decreases by 0.09 points in the same-language while it increases by 1.00 in the cross-language case). This suggests that the upper layers are *task-specific* and *language-agnostic*, since re-initializing them have minimal change on performance. We conclude that mBERT's upper layers do not contribute to cross-language transfer.

Does the Target Domain Matter? In order to test whether this behavior is specific to the cross-language setting and is not general to out-of-distribution (OOD) transfer, we repeat the same RANDOM-INIT experiment by evaluating on same-language setting while varying the evaluated domain.[3] If the drop is similar to cross-language performance, it means that lower layers are important for out-of-distribution transfer in general. Otherwise, it would confirm that these layers play a specific role for cross-language transfer.

We report the results in Table 2. For all analyzed domains (Web, News, Literature, etc.) applying RANDOM-INIT to the two first layers of the models leads to very moderate drops (e.g. -0.91 when the target domain is English Literature for parsing), while it leads to large drops when the evaluation is done on a distinct language (e.g. -5.82 when evaluated on French). The trends are similar for all the domains and tasks we tested on. We conclude that the pretrained parameters at the lower layers are consistently more critical for cross-language transfer than for same-language transfer, and cannot be explained by the possibly different domain of the evaluated datasets.

[3] Although other factors might play a part in out-of-distribution, we suspect that domains plays a crucial part in transfer. Moreover, it was shown that BERT encodes out-of-the-box domain information (Aharoni and Goldberg, 2020)

		RANDOM-INIT *of layers*				
SRC - TRG	REF	Δ0-1	Δ2-3	Δ4-5	Δ6-7	Δ8-9	Δ10-11
Domain Analyses				*Parsing*			
EN - EN	90.40	-1.41	-2.33	-1.57	-1.43	-0.60	-0.46
EN - EN LIT.	77.91	-0.91	-1.38	-1.85	-0.83	-0.23	-0.17
EN - EN WEB	75.77	-2.14	-2.42	-2.54	-1.42	-0.71	-0.69
EN - EN UGC	45.90	-1.97	-2.75	-2.10	-1.04	-0.39	-0.25
Cross-Language							
EN - FR TRAN.	83.25	-5.82	-2.69	-2.42	-0.44	0.25	0.94
EN - FR WIKI	71.29	-7.86	-4.33	-4.64	-0.92	-0.11	0.33
Domain Analyses				*POS*			
EN - EN	96.83	-1.35	-0.98	-0.70	-0.40	-0.28	-0.24
EN - EN LIT.	93.09	-0.58	-0.65	-0.28	-0.04	-0.06	0.12
EN - EN WEB	89.67	-1.07	-1.21	-0.41	-0.10	0.03	0.21
EN - EN UGC	68.93	-2.38	-1.07	-0.14	0.54	-0.04	0.63
Cross-Language							
EN - FR TRAN.	93.43	-3.59	-0.88	-1.31	-0.56	0.46	0.25
EN - FR.	91.13	-5.10	-0.93	-1.16	-0.74	0.15	-0.07
Domain Analyses				*NER*			
EN - EN	83.22	-2.45	-2.15	-1.28	-0.49	-0.15	-0.06
EN - NEWS	51.72	-1.32	-1.05	-0.80	-0.14	-0.31	-0.33
Cross-Language							
EN - FR	76.16	-5.14	-2.82	-1.97	-0.33	0.52	0.34

Table 2: Relative Zero shot Cross-Lingual performance of mBERT with RANDOM-INIT (§2.1) on pairs of consecutive layers compared to mBERT without any random-initialization (REF). We present experiments with English as the source language and evaluate across various target domains in English in comparison with the cross-lingual setting when we evaluate on French. EN-LIT. refers to the Literature Domain. UGC refers to User-Generated Content. FR-TRAN. refers to sentences translated from the English *In-Domain* test set, hence reducing the domain-gap to its minimum.
≥ REF | < REF | ≤ -2 points | ≤ -5 points

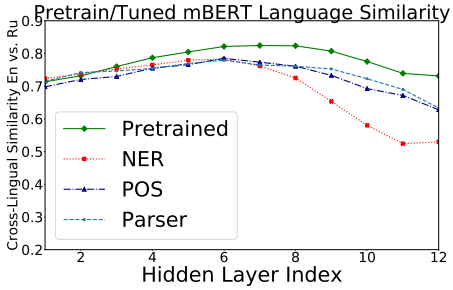


Figure 1: Cross-Lingual similarity (CKA) between representations of pretrained and fine-tuned models on POS, NER and Parsing between English and Russian.

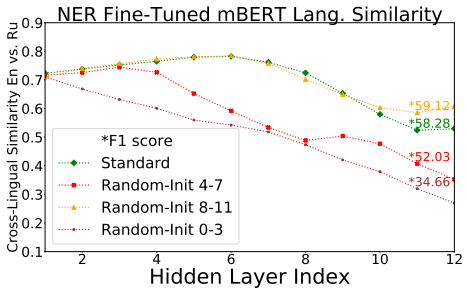


Figure 2: Cross-Lingual similarity (CKA) of the representations of a fine-tuned model on NER with and w/o RANDOM-INIT between English (source) and Russian (target). The higher the score the greater the similarity.

4.2 Cross-Lingual Similarity in mBERT

The results from the previous sections suggest that the lower layers of the model are responsible for the cross lingual transfer, whereas the upper layers are language-agnostic. In this section, we assess the transfer by directly analyzing the intermediate representations and measuring the similarities of the hidden state representations between source and target languages. We compute the CKA metric (cf. §2.2) between the source and the target representations for pretrained and fine-tuned models using parallel sentences from the PUD dataset (Zeman et al., 2017). In Figure 1, we present the similarities between Russian and English with mBERT pretrained and fine-tuned on the three tasks.[4]

The cross-lingual similarity between the representations constantly increases up to layer 5 for all the three tasks (reaching 78.1%, 78.1% and 78.2% for parsing, POS tagging and NER respectively). From these layers forward, the similarity decreases. We observe the same trends across all languages (cf. Figure 5). This demonstrates that the fine-tuned model creates similar representations regardless of the language and task, and hints on an alignment that occurs in the lower part of the model. Interestingly, the same trend is also observed in the pretrained model, suggesting that the fine-tuning step preserves the multilingual alignment.

These results do not match the findings of Singh et al. (2019), who found no language alignment across layers, although they inspected Natural Language Inference, a more "high-level task" (Dagan et al., 2005; Bowman et al., 2015). We leave the inspection of this mismatch to future work.

4.3 Better Alignment Leads to Better Cross-Lingual Transfer

In the previous section we showed that fine-tuned models align the representations between parallel sentences, across languages. Moreover, we demonstrated that the lower part of the model is critical for cross-language transfer but hardly impacts the same-language performance. In this section, we show that the alignment measured plays a critical role in cross-lingual transfer.

As seen in Figure 2 in the case of English to Russian (and in Figures 6-8 in the Appendix for other languages), when we randomly-initialize the lower part of the model, there is no alignment: the

[4] We report the comparisons for 5 other languages in Figure 5 in the Appendix.

similarity between the source and target languages decreases. We observe the same trend for all other languages and tasks and report it in the Appendix in Figures 6-8. This result matches the drop in cross-lingual performance that occurs when we apply RANDOM-INIT to the lower part of the model while impacting moderately same-language performance.

For a more systematic view of the link between the cross-lingual similarities and the cross-language transfer, we measure the Spearman correlation between the *cross-lang gap* (i.e the difference between the same-language perfromance and the cross-language performance) (Hu et al., 2020) and the cross-lingual similarity averaged over all the layers. We measure it with the cross-lingual similarity computed on the pretrained and fine-tuned models (without random-initialization) on all the languages. We find that the cross-lingual similarity correlates significantly with the *cross-lang gap* for all three tasks, both on the fine-tuned and pretrained models. The spearman correlation for the fine-tuned models are 0.76, 0.75 and 0.47 for parsing, POS and NER, respectively.[5] In summary, our results show that the cross-lingual alignment is highly correlated with the cross-lingual transfer.

5 Discussion

Understanding the behavior of pretrained language models is currently a fundamental challenge in NLP. A popular approach consists of probing the intermediate representations with external classifiers (Alain and Bengio, 2017; Adi et al., 2017; Conneau et al., 2018) to measure if a specific layer captures a given property. Using this technique, Tenney et al. (2019) showed that BERT encodes linguistic properties in the same order as the "classical NLP pipeline". However, probing techniques only indirectly explain the behavior of a model and do not explain the relationship between the information captured in the representations and its effect on the task (Elazar et al., 2020). Moreover, recent works have questioned the usage of probing as an interpretation tool (Hewitt and Liang, 2019; Ravichander et al., 2020). This motivates our approach to combine a structural analysis based on representation similarity with behavioral analysis. In this regard, our findings extend recent work from Merchant et al. (2020) in the multilingual setting, who show that fine-tuning impacts mainly the upper layers of the model and preserves the linguistic features learned during pretraining. In our case, we show that the lower layers are in charge of aligning representations across languages and that this cross-lingual alignment learned during pretraining is preserved after fine-tuning.

[5]Correlations for both the pretrained and the fine-tuned models are reported in the Appendix Table 4.

6 Conclusion

The remarkable performance of multilingual languages models in zero-shot cross-lingual transfer is still not well understood. In this work, we combine a *structural* analysis of the similarities between hidden representation across languages with a novel *behavioral* analysis that randomly-initialize the models' parameters to understand it. By combining those experiments on 17 languages and 3 tasks, we show that mBERT is constructed from: (1) a multilingual encoder in the lower layers, which aligns hidden representations across languages and is critical for cross-language transfer, and (2) a task-specific, language-agnostic predictor that has little effect to cross-language transfer, in the upper layers. Additionally, we demonstrate that hidden cross-lingual similarity strongly correlates with downstream cross-lingual performance suggesting that this alignment is at the root of these cross-lingual transfer abilities. This shows that mBERT reproduces the standard cross-lingual pipeline described by Ruder et al. (2019) without any explicit supervision signal for it. Practically speaking, our findings provide a concrete tool to measure cross-lingual representation similarity that could be used to design better multilingual pretraining processes.

Acknowledgments

We want to thank Hila Gonen, Shauli Ravfogel and Ganesh Jawahar for their insightful reviews and comments. We also thank the anonymous reviewers for their valuable suggestions. This work was partly funded by two French National funded projects granted to Inria and other partners by the Agence Nationale de la Recherche, namely projects PARSITI (ANR-16-CE33-0021) and SoSweet (ANR-15-CE38-0011), as well as by the third author's chair in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. Yanai Elazar is grateful to be partially supported by the PBC fellowship for outstanding Phd candidates in Data Science.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Chinatsu Aone and Douglas McKee. 1993. A language-independent anaphora resolution system for understanding multilingual texts. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 156–163.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Y. Goldberg. 2020. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *arXiv: Computation and Language*.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not greek to mbert: Inducing word-level translations from multilingual bert. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56.

Rob van der Goot and Gertjan van Noord. 2018. Modeling input uncertainty in neural network dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4991.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, 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, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

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.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2020. Probing the probing paradigm: Does probing accuracy entail task relevance? *arXiv preprint arXiv:2005.00719*.

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

Tanja Schultz and Alex Waibel. 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1-2):31–51.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. Bert is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 682–686. Association for Computational Linguistics.

Corso Svizzera. 2014. Converting the parallel treebank partut in universal stanford dependencies.

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.

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.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

A Appendices

A.1 Reproducibility

A.1.1 Optimization

We fine-tune our models using the standard Adam optimizer (Kingma and Ba, 2015). We warmup the learning rate on the first 10% steps and use linear decay in the rest of the training. Using the validation set of the source language, we find the best combination of hyper-parameters with a grid search on batch size among {16, 32} and learning rate initialization among {1e-5, 2.5e-5, 5e-5} We select the model with the highest validation performance out of 15 epochs for parsing and out of 6 epochs for POS tagging and NER.

Hyperparameters In Table 3, we report the best hyper-parameters set for each task, the bound of each hyperparameter, the estimated number of grid search trial for each task as well as the estimated run time.

A.1.2 Data

Data Sources We base our experiments on data originated from two sources. The Universal Dependency project (McDonald et al., 2013) downloadable here https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2988 and the WikiNER dataset (Pan et al., 2017). We also make use of the CoNLL-2003 shared task NER English dataset https://www.clips.uantwerpen.be/conll2003/

Languages For all our experiments, we use English, Russian and Arabic as source languages in addition to Chinese, Czech, Finish, French, Indonesian, Italian, Japanese, German, Hindi, Polish, Portuguese, Slovenian, Spanish, and Turkish as target languages.

Fine-tuning Data For all the cross-lingual experiments, we use English, Russian and Arabic as source languages on which we fine-tune mBERT. For English, we take the English-EWT treebank for fine-tuning, for Russian the Russian-GSD treebank and for Arabic the Arabic-PADT treebank.

Evaluation Data For all our experiments, we perform the evaluation on all the 17 languages. For Parsing and POS tagging we use the test set from the PUD treebanks released for the CoNLL Shared Task 2017 (Zeman et al., 2017). For NER, we use the corresponding annotated datasets in the wikiner dataset.

Domain Analysis Datasets We list here the datasets for completing our domain analysis experiment in Section 4.1 reported in Table 2. To have a full control on the source domains, we use for fine-tuning the English Partut treebank for POS tagging and parsing (Svizzera, 2014). It is a mix of legal, news and wikipedia text. For NER, we keep the WikiANN dataset (Pan et al., 2017). For the same-language and out-of-domain experiments, we use the English-EWT, English-Lines and English Lexnorm (van der Goot and van Noord, 2018) treebanks for Web Media data, Literature data and Noisy tweets respectively. For the cross-language French evaluation, we use the translation of the English test set,[6] as well as the French-GSD treebank. For NER, we take the CoNLL-2003 shared task English data as our out-of-domain evaluation extracted from the *News* domain. We note that the absolute performance on this dataset is not directly comparable to the one on the source wikiner. Indeed, the CoNLL-2003 dataset uses an extra MISC class. In our work, we only interpret the relative performance of different models on this test set.

Cross-Lingual Similarity Analysis For a given source language l and a target language l', we collect a 1000 pairs of aligned sentences from the UD-PUD treebanks (Zeman et al., 2017). For a given model and for each layer, we get a single sentence embedding by averaging token-level embeddings (after excluding special tokens). We then concatenate the 1000 sentence embedding vectors and get the matrices X_l and X'_l. Based on these two matrices, the CKA between the language l and the language l' is defined as:

$$CKA(X_l, X_{l'}) = \frac{||X_l^T X_{l'}||_F^2}{||X_l^T X_l||_F ||X_{l'}^T X_{l'}||_F}$$

with $||.||_F$ defining the Frobenius norm.

We do so for each source-target language pairs using the representation of the pretrained mBERT model as well as for mBERT fine-tuned on each downstream task.

In addition to the results presented in §4.2, we report in Figure 4, a comparison of the cross-lingual similarity per hidden layer of mBERT fine-tuned on NER, across target languages. The trend is the same for all languages.

[6] We do so by taking the French-ParTUT test set that overlaps with the English-ParTUT, which includes 110 sentences.

A.1.3 Computation

Infrastructure Our experiments were ran on a shared cluster on the equivalent of 15 Nvidia Tesla T4 GPUs.[7]

Codebase All of our experiments are built using the Transformers library described in (Wolf et al., 2020). We also provide code to reproduce our experiments at https://github.com/benjamin-mlr/first-align-then-predict.git.

A.1.4 Preprocessing

Our experiments are ran with word-level tokenization as provided in the datasets. We then tokenize each sequence of words at the sub-word level using the *Wordpiece* algorithm of BERT and provided by the Transformers library.

Params.	Parse	NER	POS	Bounds
batch size	32	16	16	[1,256]
learning rate	5e-5	3.5e-5	5e-5	[1e-6,1e-3]
epochs (best)	15	6	6	[1,50]
#grid	60	60	180	-
Run-time (min)	32	24	75	-

Table 3: Fine-tuning best hyper-parameters for each task as selected on the validation set of the source language with bounds. #grid: number of grid search trial. Run-time is reported in average for training and evaluation.

A.2 Cross-lingual transfer analyses

A.2.1 Correlation

We report here in Figure 4 the correlation between the hidden representation of each layer and the *cross-lang gap* between the source and the target averaged across all target languages and all layers. The correlation is strong and significant for all the tasks and for both the fine-tuned and the pretrained models. This shows that multilingual alignment that occurs within the models, learnt during pretraining is strongly related with cross-lingual transfer.

We report in Figure 3, the detail of this correlation per layer. For the pretrained model, we observe the same distribution for each task with layer 6 being the most correlated to cross-lingual transfer. We observe large variations in the fine-tuned cases, the most notable being NER. This illustrates the task-specific aspect of the relation between cross-lingual similarity and cross-lingual transfer. More precisely, in the case of NER, the sharp increase and decrease in the upper part of the model provides new evidence that for this task, fine-tuning highly impacts cross-lingual similarity in the upper part of the model which correlates with cross-language transfer.

[7] https://www.nvidia.com/en-sg/data-center/tesla-t4/

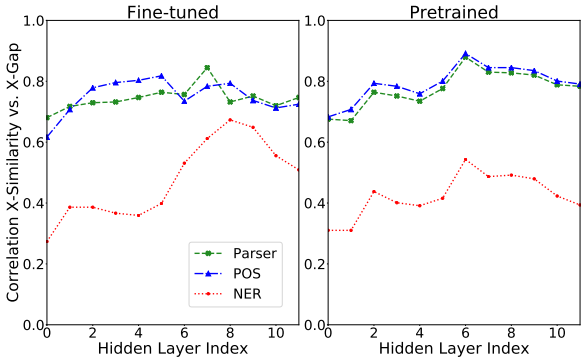


Figure 3: Spearman Correlation between Cross-Lingual Similarity (CKA between English and the target representations) and *cross-lang gap* averaged over all 17 target languages for each layer

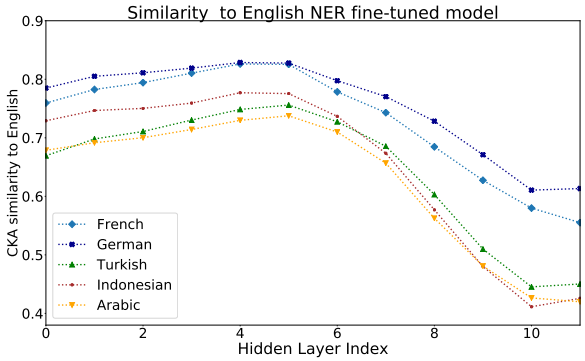


Figure 4: Cross-Lingual Similarity (CKA) (§4.2) of hidden representations of a source language (English) sentences with target languages of mBERT fine-tuned for NER. The higher the CKA value the greater the similarity.

Task	Correlation X-Gap vs. X-Similarity	
	FINE-TUNED	PRETRAINED
Parsing	0.76	0.79
POS	0.74	0.82
NER*	0.47	0.43

Table 4: Spearman-Rank Correlation between the *Cross-lingual Gap* (X-Lang Gap) and the *Cross-lingual Similarity* between the source and the target languages of the fine-tuned models and the pretrained model averaged over all the hidden layers and all the 17 target languages (sample size per task: 17). For NER, *cross-lang gap* measured on wikiner data and not on the parrallel data itself in constrast with Parsing and POS tagging. Complete list of languages can be found in Appendix A.1.2

Eval	REF	ALL	RANDOM-INIT *of layers* 1	2	1-2	3-4	1-3	4-6	7-9	10-12	1-4	5-8	9-12
						Parsing							
ENGLISH DEV	88.52	74.66	87.77	88.03	87.28	86.81	83.77	85.86	87.53	88.78	84.30	85.41	88.35
ENGLISH TEST	88.59	74.58	87.77	88.09	87.25	86.79	83.37	85.54	87.36	88.62	83.10	85.37	88.69
FRENCH	68.94	3.70	65.73	65.21	55.31	61.31	43.81	61.77	67.03	69.36	37.29	61.82	69.26
GERMAN	67.43	4.73	64.97	65.20	57.08	60.62	47.85	58.93	64.12	66.67	36.05	59.37	67.21
TURKISH	28.40	2.76	21.65	23.77	16.78	21.21	10.69	20.23	25.39	30.43	9.70	20.94	29.33
INDONESIAN	45.13	4.99	43.33	43.48	39.83	39.09	33.06	40.65	44.42	46.96	30.35	40.85	47.53
RUSSIAN	59.70	2.95	57.81	57.53	54.10	53.51	47.01	52.37	56.45	61.41	38.58	52.41	60.72
ARABIC	23.37	3.19	23.66	23.49	21.01	19.55	16.17	18.84	20.70	24.54	13.26	18.27	23.93
						POS							
ENGLISH DEV	96.45	87.47	96.04	96.06	95.92	95.81	95.38	95.43	96.25	96.58	94.01	95.35	96.39
ENGLISH TEST	96.53	87.71	96.08	96.24	95.94	95.72	95.40	95.59	96.34	96.74	94.05	95.45	96.51
FRENCH	88.25	28.96	86.70	87.66	79.84	87.14	69.43	86.42	86.94	88.30	62.28	86.37	88.26
GERMAN	90.63	28.93	88.26	89.53	82.26	88.39	71.63	88.30	90.26	90.83	59.16	89.12	90.64
TURKISH	72.65	32.23	62.17	66.17	54.50	63.22	47.77	66.37	70.91	72.92	44.16	69.30	73.08
INDONESIAN	84.06	36.98	82.15	82.89	80.13	81.40	75.94	81.99	83.78	84.42	72.42	82.59	84.09
RUSSIAN	82.97	32.63	83.14	83.63	81.95	82.26	77.93	81.69	82.98	81.76	70.33	82.56	83.19
ARABIC	56.66	19.61	58.10	58.06	57.89	55.62	57.93	54.69	56.04	55.97	52.28	53.60	58.84
						NER							
ENGLISH DEV	83.29	56.99	82.04	82.26	79.52	80.36	76.22	79.53	82.18	82.53	69.31	80.05	82.47
ENGLISH TEST	83.06	56.56	81.46	82.00	79.63	79.25	76.68	78.93	81.64	82.39	69.08	79.91	82.27
FRENCH	76.76	35.35	75.46	77.57	69.94	72.83	65.14	70.34	75.42	75.90	55.79	73.12	75.77
GERMAN	76.68	18.95	73.73	75.39	66.18	70.12	56.50	69.53	75.38	77.11	42.37	71.14	75.50
TURKISH	67.64	20.76	62.54	64.84	52.20	57.11	53.03	60.59	65.66	64.87	39.38	61.43	66.62
INDONESIAN	53.47	21.20	49.19	49.27	46.50	46.87	43.75	47.83	54.39	48.71	36.11	46.06	48.23
RUSSIAN	58.23	7.43	55.63	58.08	50.67	52.89	42.83	46.13	53.38	58.09	34.66	52.03	59.12
ARABIC	41.81	5.49	35.79	34.80	32.37	32.31	26.21	38.88	38.55	40.83	21.85	38.67	41.23

Table 5: Zero-shot cross-lingual performance when applying RANDOM-INIT to specific set of consecutive layers compared to the REF model. Source language is English. Baseline model ALL (for all layers randomly initialized) corresponds to a model trained from scratch on the task. For reproducibility purposes, we report performance on the Validation set ENGLISH DEV. For all target languages, we report the scores on the test split of each dataset. Each score is the average of 5 runs with different random seeds. For more insights into the variability of our results, we report the min., median and max. value of the standard deviations (std) across runs with different random seeds for each task: Parsing:0.02/0.34/1.48, POS:0.01/0.5/2.38, NER:0.0/0.47/2.62 (std min/median/max).

$\geq$ REF | $<$ REF | $\leq$ 5 points | $\leq$ 10 points

		RANDOM-INIT *of layers*					
SOURCE - TARGET	REF	Δ 0-1	Δ 2-3	Δ 4-5	Δ 6-7	Δ 8-9	Δ 10-11
		Parsing					
EN - ENGLISH	88.98	-0.96	-0.66	-0.93	-0.55	0.04	-0.09
EN - ARABIC	35.88	-4.05	-2.38	-3.16	-0.78	1.74	1.68
EN - FRENCH	74.04	-21.30	-6.84	-2.93	-0.69	0.03	0.76
EN - GERMAN	70.34	-15.06	-9.26	-4.75	-1.54	-0.29	1.82
EN - TURKISH	34.03	-16.37	-10.10	-5.11	-3.71	0.43	1.43
EN - INDO	44.11	-10.57	-5.87	-2.66	-0.96	-0.74	0.73
EN - RUSSIAN	62.52	-7.31	-5.37	-2.84	-1.09	0.44	0.71
EN - PORTHUGHESE	68.59	-25.83	-6.22	-2.97	-0.77	0.15	0.82
EN - SPANISH	69.96	-18.05	-5.74	-2.78	-0.96	0.13	0.72
EN - FINISH	48.42	-24.25	-9.48	-4.39	-2.51	-0.28	0.22
EN - ITALIAN	74.54	-30.54	-9.63	-4.18	-1.32	-0.12	0.90
EN - SLOVENIAN	73.04	-29.89	-6.52	-3.00	-1.68	-0.05	0.18
EN - CZECH	60.44	-31.84	-10.69	-4.61	-1.82	0.18	1.17
EN - POLISH	55.23	-23.57	-9.11	-3.34	-1.83	0.28	0.89
EN - HINDI	28.86	-9.13	-7.58	-5.84	-2.50	1.35	1.49
EN - CHINESE	27.48	-7.31	-4.47	-1.65	-0.62	0.65	1.32
EN - JAPANESE	11.99	-4.36	-2.76	-1.91	-1.19	0.47	1.12
EN - X (MEAN)	53.23	-15.77	-6.51	-3.39	-1.47	0.29	1.00
RU - RUSSIAN	85.15	-0.82	-1.38	-1.51	-0.86	-0.29	0.18
RU - ENGLISH	61.40	-8.37	-3.55	-3.90	-0.72	1.77	1.14
RU - ARABIC	59.41	-5.65	-5.26	-5.15	-1.47	0.24	0.16
RU - FRENCH	65.84	-8.87	-2.93	-1.81	-1.05	3.81	1.24
RU - GERMAN	65.90	-7.02	-4.19	-1.97	-1.45	2.58	2.05
RU - TURKISH	32.20	-13.13	-7.18	-6.82	-3.77	-0.85	1.21
RU - INDO	47.59	-4.74	-2.99	-2.30	-1.81	0.04	1.02
RU - PORTHUGHESE	66.41	-11.17	-1.61	-1.09	-1.25	4.16	1.94
RU - SPANISH	66.74	-4.52	-1.38	-0.69	-0.97	2.95	1.37
RU - FINISH	52.92	-15.43	-6.59	-4.09	-1.35	0.12	0.77
RU - ITALIAN	65.28	-12.97	-3.56	-2.34	-1.46	3.16	1.55
RU - SLOVENIAN	62.91	-16.67	-2.71	-3.18	-1.03	0.31	1.08
RU - CZECH	72.77	-11.95	-4.17	-3.13	-1.57	-0.33	0.30
RU - POLISH	66.07	-5.70	-3.22	-2.57	-1.54	-0.12	0.54
RU - HINDI	28.67	-6.02	-5.77	-5.27	-3.75	-0.06	0.99
RU - CHINESE	28.77	-4.66	-4.38	-3.22	-1.80	0.15	1.12
RU - JAPANESE	15.10	-4.89	-3.56	-3.95	-3.11	0.68	0.73
RU - X (MEAN)	55.41	-7.69	-3.71	-3.13	-1.70	0.92	0.94
AR - ARABIC	59.54	-0.78	-2.14	-1.20	-0.67	-0.27	0.08
AR - ENGLISH	25.46	-2.09	-2.92	-0.90	-1.40	-0.97	-0.61
AR - FRENCH	28.92	-4.85	-1.45	-0.25	-2.72	-1.60	-0.88
AR - GERMAN	27.14	-6.38	-4.51	-0.98	-2.24	0.13	0.09
AR - TURKISH	9.58	-3.90	-3.14	-2.76	-2.33	0.31	0.15
AR - INDO	36.16	-5.85	-4.86	-1.71	-0.68	-0.17	0.58
AR - RUSSIAN	42.25	-3.52	-5.28	-2.46	-1.66	-0.67	-0.27
AR - PORTHUGHESE	34.71	-4.80	-1.22	0.10	-2.98	-0.33	-0.24
AR - SPANISH	31.95	-4.02	-0.15	-0.44	-1.46	-0.77	0.38
AR - FINISH	28.18	-9.89	-7.03	-3.17	-1.81	-0.58	-0.42
AR - ITALIAN	28.85	-3.01	0.60	1.45	-2.26	-1.47	-0.70
AR - SLOVENIAN	35.78	-9.73	-4.97	-2.21	-1.43	-0.41	-0.56
AR - CZECH	40.04	-13.61	-6.82	-3.20	-2.38	-1.12	-0.21
AR - POLISH	41.16	-8.46	-5.52	-2.48	-1.48	-0.47	-0.55
AR - HINDI	10.24	-2.46	-2.86	-2.57	-1.55	1.00	0.14
AR - CHINESE	11.46	-2.42	-2.43	-1.26	-0.82	0.23	-0.05
AR - JAPANESE	6.66	-1.28	-0.79	-1.20	-1.04	0.74	0.30
AR - X (MEAN)	27.97	-4.91	-3.17	-1.48	-1.68	-0.36	-0.14

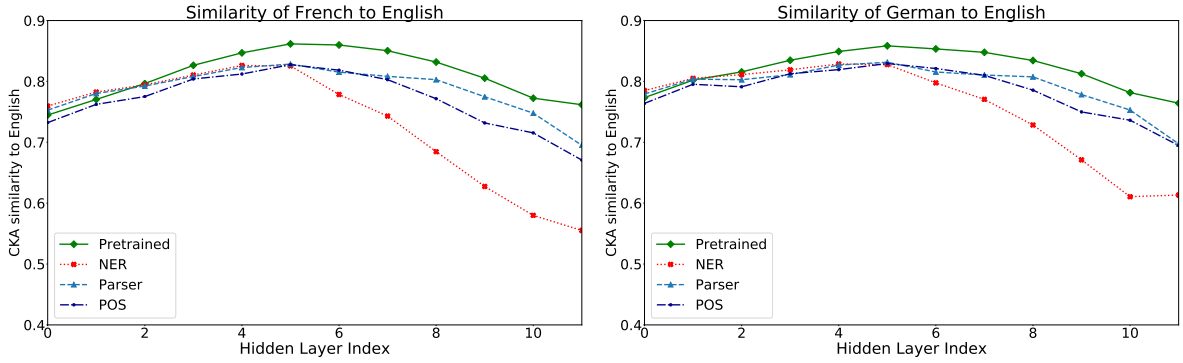
Table 6: Parsing (LAS score) Relative Zero shot Cross-Lingual performance of mBERT with RANDOM-INIT (section 2.1) on pairs of consecutive layers compared to mBERT without any random-initialization (REF). In SRC - TRG, SRC indicates the source language on which we fine-tune mBERT, and TRG the target language on which we evaluate it. SRC-X is the average across all 17 target language with X $\neq$ SRC $\geq$ REF $<$ REF $\leq$ -2 points $\leq$ -5 points

Source - Target	Ref	Random-init *of layers* Δ 0-1	Δ 2-3	Δ 4-5	Δ 6-7	Δ 8-9	Δ 10-11
				POS			
En - English	96.51	-0.30	-0.25	-0.40	-0.00	0.05	0.02
En - Arabic	70.20	-3.63	-1.88	-2.40	-1.26	-1.89	-2.74
En - French	89.16	-9.68	-2.09	-1.49	-1.03	0.29	0.59
En - German	89.32	-7.81	-2.12	-1.27	-0.99	-0.46	-0.68
En - Turkish	71.67	-11.62	-4.43	-1.48	-0.95	0.04	-0.95
En - Indo	71.44	-6.39	-2.80	-1.74	-0.59	-0.41	-1.10
En - Russian	86.26	-2.66	-0.94	-0.27	0.13	0.37	0.62
En - Porthughese	86.51	-10.84	-1.83	-1.44	-0.81	-0.01	-0.14
En - Spanish	87.26	-8.09	-1.30	-1.36	-1.13	0.20	0.17
En - Finish	84.85	-20.00	-8.09	-2.77	-0.97	-0.06	-0.86
En - Italian	91.35	-13.97	-3.35	-2.66	-1.34	-0.01	0.27
En - Slovenian	89.64	-16.46	-2.41	-1.09	-0.18	0.34	0.19
En - Czech	83.39	-19.62	-3.93	-0.73	-0.56	0.21	0.29
En - Polish	81.45	-13.33	-3.52	-1.19	-1.22	-0.50	-0.16
En - Hindi	65.43	-10.04	-2.70	-2.89	-3.25	3.00	0.28
En - Chinese	67.89	-3.04	-2.82	-3.59	-0.29	0.66	0.29
En - Japanese	48.86	-2.19	1.52	-1.51	-1.13	1.42	1.79
En - X (Mean)	79.37	-8.94	-2.49	-1.66	-0.88	0.20	-0.14
Ru - Russian	96.90	-0.52	-0.55	-0.40	-0.07	0.02	-0.03
Ru - English	82.55	-20.72	-7.06	-5.01	-3.93	0.74	-1.57
Ru - Arabic	79.30	-4.04	-1.48	-2.06	0.64	0.01	0.47
Ru - French	86.02	-18.66	-4.64	-4.10	-9.00	-0.13	-1.84
Ru - German	84.90	-12.50	-4.80	-2.79	-3.90	0.47	-1.82
Ru - Turkish	69.92	-15.20	-2.06	-0.55	-1.41	-0.11	0.68
Ru - Indo	71.16	-8.33	-3.44	-1.03	-0.56	-0.73	0.15
Ru - Porthughese	84.24	-19.56	-7.15	-3.00	-7.78	-0.15	-2.08
Ru - SPanish	84.84	-13.64	-4.09	-2.66	-7.67	-0.35	-2.48
Ru - Finish	81.08	-18.55	-5.42	-1.37	-1.00	-0.16	0.02
Ru - Italian	85.56	-21.04	-5.11	-3.41	-8.21	-0.20	-3.36
Ru - Slovenian	85.37	-14.65	-3.53	-1.72	-2.00	-0.15	-0.15
Ru - Czech	87.37	-8.43	-1.99	-0.71	-1.16	-0.50	-0.28
Ru - Polish	86.42	-4.41	-1.89	-0.64	-0.44	-0.21	0.09
Ru - Hindi	65.49	-1.16	0.41	-1.49	-2.17	1.13	3.20
Ru - Chinese	65.85	-5.12	-1.43	-0.32	-0.74	-0.13	-0.47
Ru - Japanese	46.91	-0.72	2.16	0.00	-1.30	1.15	1.12
Ru - X (Mean)	79.25	-10.08	-2.83	-1.65	-2.74	0.01	-0.45
Ar - Arabic	79.28	-0.35	-0.49	-0.36	-0.19	-0.05	-0.00
Ar - English	63.26	-3.32	-1.09	-1.72	-1.68	-1.03	-1.78
Ar - French	63.33	-4.41	-1.53	-1.14	-1.30	-0.44	-0.92
Ar - German	63.23	-4.95	-2.97	-1.04	-1.58	-0.53	-2.09
Ar - Turkish	60.99	-13.76	-8.74	-2.86	-4.49	-1.08	-1.88
Ar - Indo	64.24	-5.11	-3.43	-1.87	-0.58	-0.28	-0.63
Ar - Russian	74.52	-4.01	-2.37	-2.40	-1.84	-1.69	-2.03
Ar - Porthughese	67.28	-6.51	-2.84	-1.30	-1.23	0.04	-0.96
Ar - SPanish	64.84	-3.08	-0.51	-0.74	-0.48	0.02	-0.14
Ar - Finish	64.28	-19.72	-8.32	-3.72	-2.56	-1.64	-3.03
Ar - Italian	63.55	-4.25	-1.60	-0.94	-1.15	0.14	-0.64
Ar - Slovenian	68.06	-12.21	-4.31	-2.17	-1.85	0.68	-1.81
Ar - Czech	72.65	-13.57	-3.14	-1.88	-1.77	-1.35	-1.57
Ar - Polish	75.00	-8.87	-2.94	-1.46	-0.62	-1.00	-1.37
Ar - Hindi	62.29	-7.31	-6.07	-2.42	-1.26	0.19	-1.72
Ar - Chinese	56.51	-5.02	-4.94	-2.10	-1.35	-1.02	-1.77
Ar - Japanese	47.06	-3.34	-3.34	-0.65	-0.89	-1.54	-0.35
Ar - X (Mean)	64.81	-6.73	-3.50	-1.63	-1.56	-0.73	−1.29

Table 7: POS tagging Relative Zero shot Cross-Lingual performance of mBERT with Random-init (section 2.1) on pairs of consecutive layers compared to mBERT without any random-initialization (Ref). In Src - Trg, Src indicates the source language on which we fine-tune mBERT, and Trg the target language on which we evaluate it. Src-X is the average across all 17 target language with $X \neq$ Src. $\geq$ Ref $<$ Ref $\leq$ -2 points $\leq$ -5 points

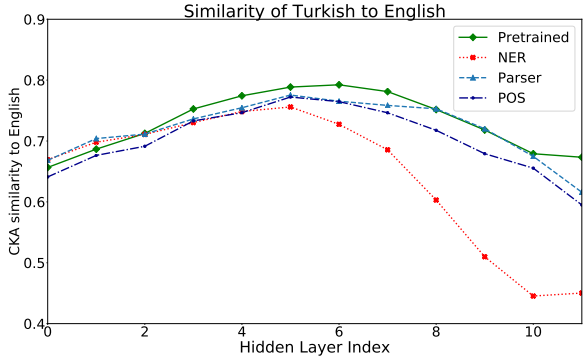
Source - Target	Ref	Random-init *of layers* Δ 0-1	Δ 2-3	Δ 4-5	Δ 6-7	Δ 8-9	Δ 10-11
				NER			
EN - English	83.27	-2.64	-2.12	-1.41	-0.61	-0.21	-0.14
EN - French	76.20	-4.41	-2.72	-2.09	-0.30	0.51	0.08
EN - German	75.58	-8.25	-4.65	-2.50	-0.40	0.06	0.26
EN - Turkish	66.23	-8.71	-6.57	-2.16	-1.01	0.51	0.51
EN - Indo	50.24	-2.94	-1.43	-2.54	2.49	-0.70	0.82
EN - Porthughese	76.09	-4.66	-0.88	-1.16	-0.57	0.62	-0.70
EN - SPanish	67.00	-0.99	4.37	2.03	-1.69	1.57	-1.38
EN - Finish	75.61	-11.89	-4.47	-2.29	0.63	0.54	-0.37
EN - Italian	78.48	-6.65	-3.64	-3.08	-1.32	-0.30	-0.28
EN - Slovenian	72.80	-10.37	-2.96	-3.11	-0.36	0.10	-0.72
EN - Czech	76.90	-8.02	-6.81	-3.17	0.09	1.00	0.39
EN - Russian	60.20	-5.87	-6.65	-5.71	-2.82	-0.82	-0.37
EN - Arabic	39.15	-8.98	-5.31	-1.97	1.56	0.31	-0.98
EN - Polish	77.20	-8.32	-5.53	-3.05	-0.06	0.67	0.09
EN - Hindi	60.61	-12.08	-13.88	-9.23	-0.91	-1.25	2.08
EN - Chinese	37.74	-13.68	-6.49	-4.59	-2.41	-5.23	-1.00
EN - Japanese	25.19	-11.40	-7.54	-4.67	-2.53	-3.45	-0.23
EN - X (mean)	64.17	-8.28	-5.09	-3.07	-0.79	-0.47	-0.13
Ru - Russian	88.20	-2.08	-2.13	-1.52	-0.64	-0.33	-0.13
Ru - English	56.62	-13.83	-8.52	-4.70	-1.50	-0.76	1.38
Ru - French	67.35	-18.45	-9.70	-4.32	-1.76	-1.77	2.29
Ru - German	69.23	-13.94	-9.01	-5.80	-2.98	-1.65	0.40
Ru - Turkish	63.64	-18.52	-10.06	-6.01	-4.16	-0.67	-0.27
Ru - Indo	41.92	-10.29	-7.20	-5.19	-1.20	-1.91	0.50
Ru - Porthughese	67.33	-21.23	-8.27	-8.84	-2.83	-1.83	1.51
Ru - SPanish	69.15	-16.74	-10.00	-8.16	-5.80	-1.66	0.26
Ru - Finish	73.03	-17.17	-8.70	-5.88	-2.12	0.86	1.48
Ru - Italian	70.05	-19.47	-9.54	-6.90	-3.06	0.73	1.04
Ru - Slovenian	71.18	-12.02	-9.48	-3.61	-0.70	1.16	2.14
Ru - Czech	74.87	-17.93	-10.59	-6.34	-4.02	0.17	-0.23
Ru - Arabic	38.63	-8.67	-6.81	-0.13	-0.65	-1.34	-0.29
Ru - Polish	75.16	-15.38	-7.97	-6.33	-3.07	-0.63	1.34
Ru - Hindi	58.01	-19.60	-12.36	-6.18	0.93	-1.64	1.17
Ru - Chinese	43.86	-23.73	-11.68	-6.80	-4.27	-4.13	-6.01
Ru - Japanese	30.79	-16.80	-11.29	-5.26	-2.77	-3.99	-6.91
Ru - X (mean)	62.13	-15.85	-9.36	-5.50	-2.44	-1.16	-0.06
Ar - Arabic	87.97	-2.37	-2.11	-0.96	-0.39	-0.15	0.21
Ar - French	75.21	-18.71	-8.31	-3.76	-0.19	0.82	1.07
Ar - German	74.24	-15.25	-7.19	-3.72	-1.38	-0.04	0.27
Ar - Turkish	68.45	-14.89	-8.65	-2.78	-0.30	0.98	1.90
Ar - Indo	54.65	-13.86	-10.95	-8.53	-4.66	-2.82	0.09
Ar - Porthughese	74.67	-20.42	-10.54	-3.17	-1.59	0.10	1.28
Ar - SPanish	74.88	-18.16	-12.18	-3.06	-1.95	0.52	0.63
Ar - Finish	78.01	-18.79	-8.84	-4.30	-2.03	-0.30	0.19
Ar - Italian	75.76	-16.37	-7.73	-3.98	-1.49	-0.06	0.74
Ar - Slovenian	63.08	-11.13	-5.49	4.79	0.88	2.17	0.79
Ar - Czech	74.70	-21.93	-10.95	-5.84	-2.42	-1.36	0.09
Ar - Russian	45.51	-7.59	-5.81	-2.63	0.15	-0.22	0.47
Ar - English	57.94	-12.79	-6.03	-4.57	-0.32	0.29	1.65
Ar - Polish	77.29	-20.61	-9.47	-5.93	-2.64	-1.09	-0.19
Ar - Hindi	65.31	-14.95	-9.12	-3.84	-1.48	0.72	0.98
Ar - Chinese	45.88	-25.72	-10.67	-3.99	-1.41	-2.72	0.57
Ar - Japanese	24.75	-14.66	-5.19	-3.82	-0.99	-1.17	1.50
Ar - X (mean)	65.59	-16.10	-8.42	-3.73	-1.40	-0.25	0.67

Table 8: NER (F1 score) Relative Zero shot Cross-Lingual performance of mBERT with Random-init (section 2.1) on pairs of consecutive layers compared to mBERT without any random-initialization (Ref). In Src - Trg, Src indicates the source language on which we fine-tune mBERT, and Trg the target language on which we evaluate it. Src-X is the average across all 17 target language with X $\neq$ Src $\geq$ Ref $<$ Ref $\leq$ -2 points $\leq$ -5 points

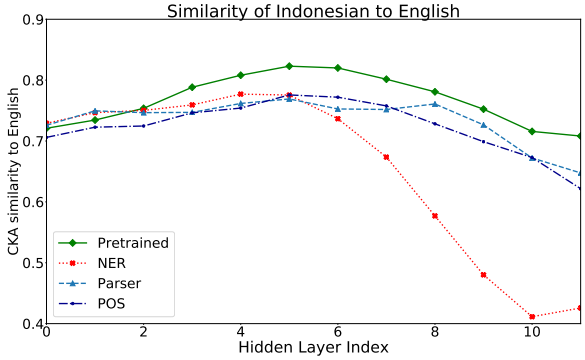


(a)

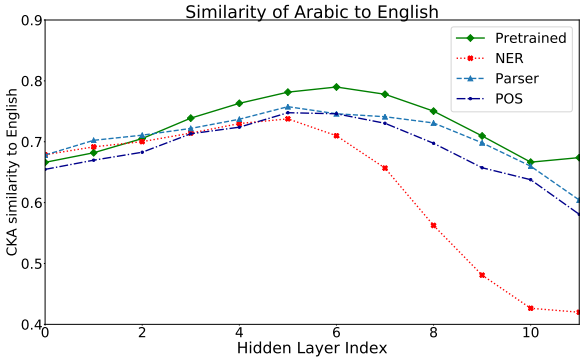
(b)



(c)

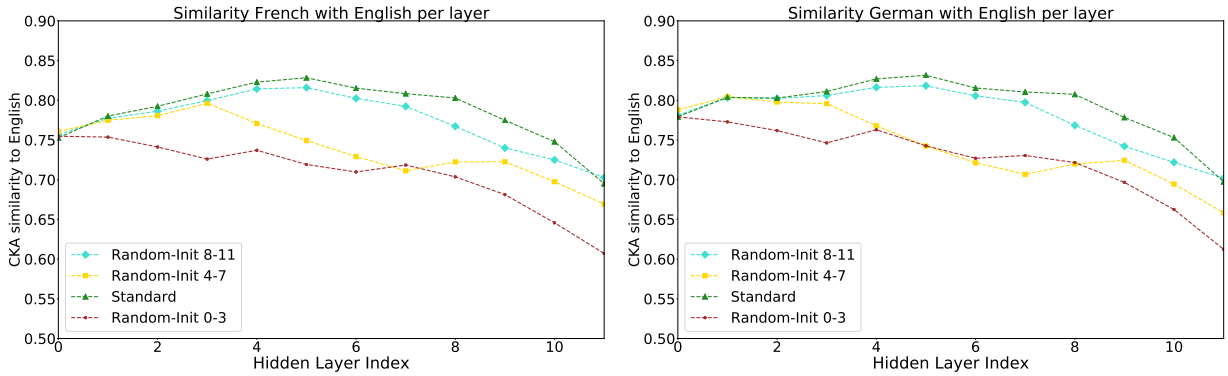


(d)



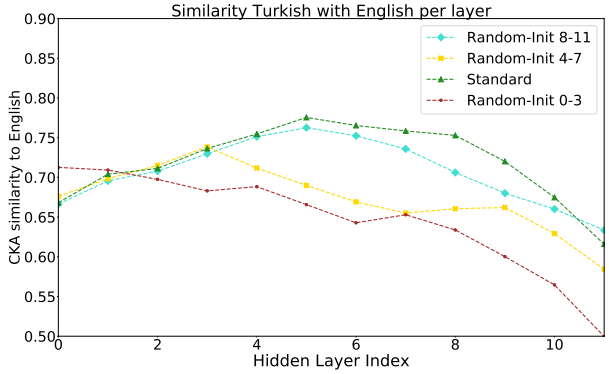
(e)

Figure 5: Cross-Lingual similarity (CKA) similarity (§4.2) of hidden representations of a source language (English) sentences with a target language sentences on fine-tuned and pretrained mBERT. The higher the CKA value the greater the similarity.

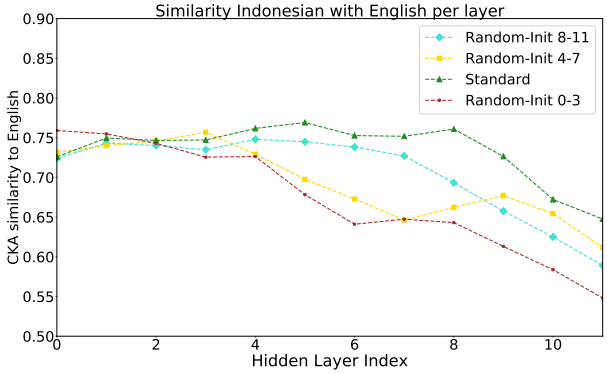


(a)

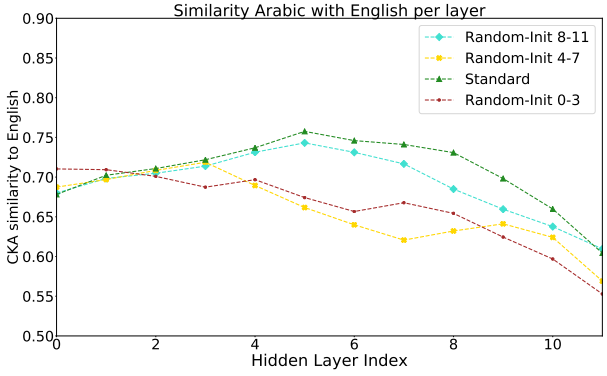
(b)



(c)



(d)



(e)

Figure 6: Cross-Lingual similarity (CKA) (§4.2) of hidden representations of a source language (English) sentences with target languages sentences on fine-tuned **Parsing** models with and without RANDOM-INIT. The higher the CKA value the greater the similarity.

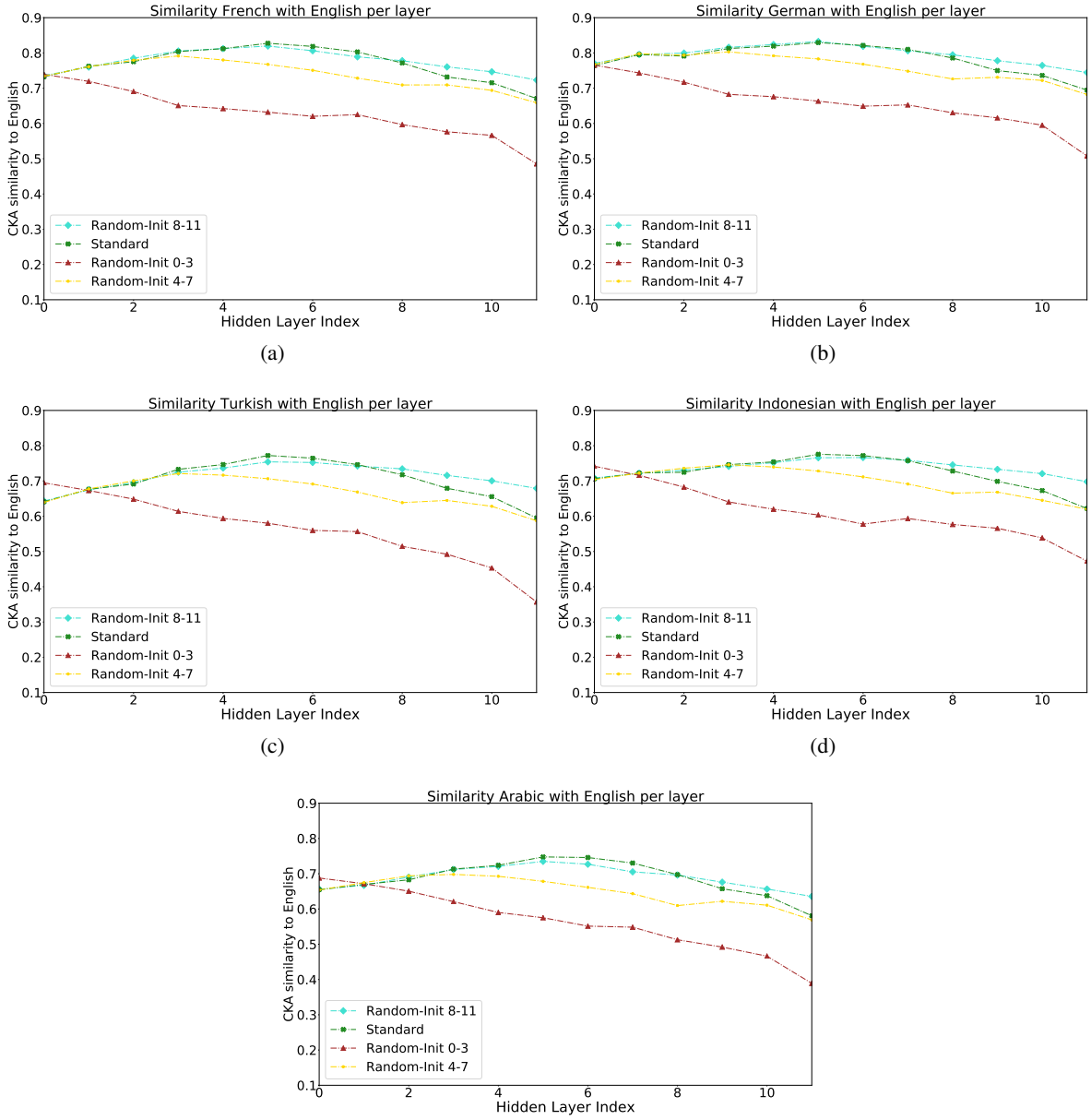
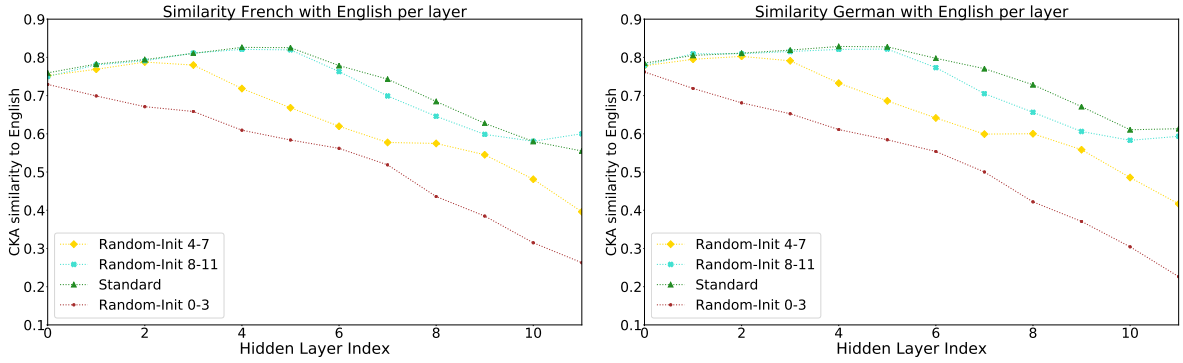
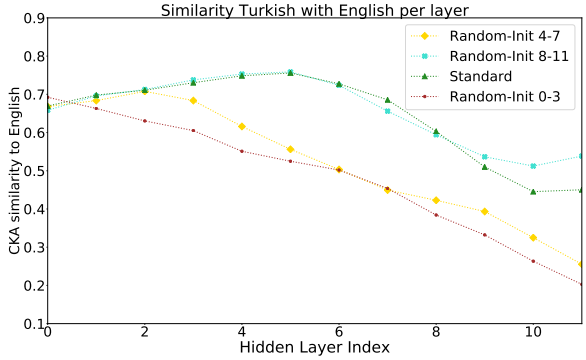


Figure 7: Cross-Lingual similarity (CKA) (§4.2) of hidden representations of a source language (English) sentences with target languages sentences on fine-tuned **POS** models with and w/o RANDOM-INIT. The higher the CKA value the greater the similarity.

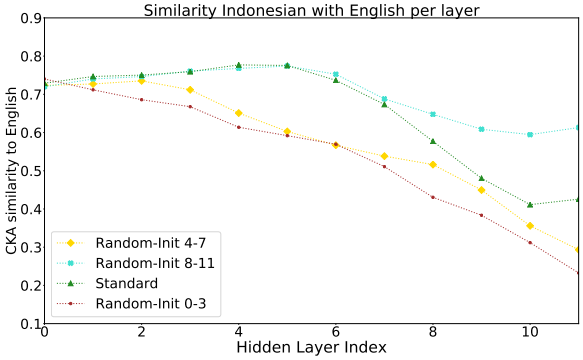


(a)

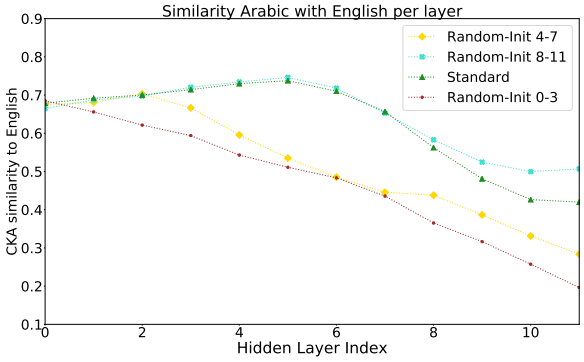
(b)



(c)



(d)



(e)

Figure 8: Cross-Lingual similarity (CKA) (§4.2) of hidden representations of a source language (English) sentences with target languages sentences on fine-tuned **NER** models with and w/o RANDOM-INIT. The higher the CKA value the greater the similarity.