BERT Cannot Align Characters

Antonis Maronikolakis and **Philipp Dufter** and **Hinrich Schütze**
Center for Information and Language Processing (CIS), LMU Munich, Germany
`antmarakis,philipp@cis.lmu.de`

Abstract

In previous work, it has been shown that BERT can adequately align cross-lingual sentences on the word level. Here we investigate whether BERT can also operate as a char-level aligner. The languages examined are English, Fake-English, German and Greek. We show that the closer two languages are, the better BERT can align them on the character level. BERT indeed works well in English to Fake-English alignment, but this does not generalize to natural languages to the same extent. Nevertheless, the proximity of two languages does seem to be a factor. English is more related to German than to Greek and this is reflected in how well BERT aligns them; English to German is better than English to Greek. We examine multiple setups and show that the similarity matrices for natural languages show weaker relations the further apart two languages are.

1 Introduction

For the many sweeping successes BERT has had in the field of Natural Language Processing, the model's alignment capabilities have been lacking and under-explored. Work in this area is picking up and it has been shown that BERT can operate with adequate efficiency in word alignment tasks (Zenkel et al., 2019; Jalili Sabet et al., 2020). The question whether or not BERT can perform character-level alignment, though, has not been answered yet.

Even though characters on their own do not necessarily hold much semantic meaning, we investigate whether BERT is able to generate useful representation spaces for characters. Character-level alignment would be useful in tasks like transliteration (Li et al., 2009; Sajjad et al., 2017) or word-level alignment (Legrand et al., 2016). In a lot of occurrences of transliterations, grammatical inflections are added or dropped, which causes difficulties in an array of tasks (Czarnowska et al.,

a	a	0.96		a	ρ	0.98
b	z	0.94		b	ξ	0.98
c	l	0.95		c	ι	0.96
d	x	0.92		d	σ	0.96
e	h	0.96		e	o	0.98
f	**f**	0.94		f	ι	0.98
g	**g**	0.97		g	λ	0.98
h	x	0.93		h	τ	0.98
i	**i**	0.97		i	τ	0.98
j	z	0.97		j	η	0.96
k	c	0.95		k	ν	0.97
l	z	0.96		l	ν	0.98
m	**m**	0.96		m	ψ	0.96
n	**n**	0.98		n	η	0.95
o	**o**	0.98		o	α	0.97
p	**p**	0.97		**p**	**π**	0.97
q	c	0.96		q	ρ	0.97
r	**r**	0.98		r	λ	0.99
s	x	0.94		s	ν	0.99
t	k	0.93		**t**	**τ**	0.98
u	z	0.93		u	φ	0.97
v	h	0.94		v	π	0.99
w	**w**	0.96		w	τ	0.98
x	**x**	0.98		x	ν	0.99
y	c	0.97		y	ν	0.98
z	c	0.94		z	τ	0.98

(a) *EngDeu*	(b) *EngEll*

Table 1: Max alignments for English → German and English → Greek, as extracted from a cosine similarity matrix between the two alphabets for each experiment. Correct alignments are in bold. We see low accuracy for both setups, but especially low for English → Greek.

2019; Vania and Lopez, 2017). With character-level awareness, we can have better models for transliteration detection and extraction tasks. Word alignment could also benefit from character-level information in instances where words get split up within a sentence (eg., separable verbs in German or phrasal verbs in English).

With our work we show that even though BERT is not able to align languages on the character level, the closer these languages are the better the alignment. In the trivial case of English to Fake-English alignment, the model successfully learns to align characters. For English to German performance drops substantially and it drops even more for English to Greek. Languages seem to be put on an intuitive scale by BERT, with more similar languages having better alignment than highly dissimilar ones.

2 Related Work

Research has been conducted to uncover elements of multilinguality in mBERT. In Pires et al. (2019), an analysis of mBERT is presented, while in Conneau and Lample (2019); Wu and Dredze (2019); Artetxe et al. (2020) zero-shot cross-lingual transfer is analyzed. Dufter and Schütze (2020) further analyzes mBERT's capabilities with BERT's architecture and the structure of languages examined. The authors performed their experiments on a pairing of English with Fake-English, as proposed by K et al. (2020) in their rigorous empirical study of mBERT where linguistic properties of languages, architecture and learning objectives are investigated.

In Wu et al. (2021); Garg et al. (2019), it is shown that transformers (Vaswani et al., 2017) can achieve similar performance with sequence-to-sequence approaches based on Recurrent Neural Networks (Luong et al., 2015) for character-level tasks such as transliteration and grapheme-to-phoneme conversion. Further work to develop character-level BERT-based models is conducted in El Boukkouri et al. (2020); Ma et al. (2020).

Och and Ney (2003); Legrand et al. (2016) worked towards representation-based word alignment, with implementations of aligners proposed in Dyer et al. (2013); Östling and Tiedemann (2016).

We also note recent efforts towards unsupervised word alignment. In Zenkel et al. (2019), an extension to the usual Machine Translation encoder-decoder is proposed to jointly learn language translation and word alignment in an unsupervised manner. BERT has also been shown to be able to perform word alignment (Jalili Sabet et al., 2020) through embedding matrix similarities.

3 Experimental Setup

3.1 Data Setup

The EuroParl Corpus (Koehn, 2005) is a parallel corpus containing recorded proceedings of the European Parliament. Originally 21 languages were included, although here we examined three: English (ENG), German (DEU) and Greek (ELL).[1] Each set is split by words and each word is further split in characters. Finally, special start and end tokens are added around each (split) word. This process results in a data file where each line contains a word split in characters. Finally, these language

---

sets are merged together, alternating between lines. For example, in the English → German setup, the first line contains an English word (split in characters), the second line a German word (split in characters), and so on. For the conversion of English to Fake-English, we employ a mapping of characters to integers. The integers are in the range $[100, 151]$. The same mapping takes place for the English → German setup, since the two languages share the same script. Data is given as input line-by-line to the model.

3.2 Model Setup

We experimented with different BERT (Devlin et al., 2019) model sizes and parameters. In our hyperparameter search, we mainly examined the effect of hidden layer size, layer numbers, embedding space size and attention heads number. We found that when models have fewer than 3 layers or more than 9 layers, we see underfitting and overfitting respectively. In the end, we settled for a 6-layer model with a quarter of the original BERT-base parameters, trained for 50 epochs. An analysis of model size effect on performance is omitted. We are training from scratch on the usual Masked Language Modeling task as described in (Devlin et al., 2019), with the difference that we are masking individual characters instead of subword tokens.

3.3 Experiments

For our control experiment, we tried to align English and Fake-English. This setup serves as aid to hyperparameter tuning. Because of the nature of English and Fake-English, if a model cannot align these two then it would not work for natural languages. English data was split in two sets, with the second converted to Fake-English. In our setup, Fake-English is a simple mapping from English characters to numbers ranging from 100 to 151. That is, 'a' is converted to '100', 'b' to '101', 'A' to '126' etc. All other numbers were removed from both sets. We call this setup $EngFake_{base}$.

Apart from the base setup, we tried minor alterations. Namely, we tried to break the one-to-one mapping from English to Fake-English. The letter 'f' was mapped not to a single, unique integer, but instead two new indices (denoted by f_1 and f_2). So, 'f' was mapped to '200 201', which are the unique tokens for f_1 and f_2 respectively. In the same way, capital 'F' was replaced by the tokens for F_1 and F_2. We call this setup $EngFake_{f_1 f_2}$.

After the successful $EngFake$ experiments, we experiment with English to Greek ($EngEll$) and English to German ($EngDeu$). For English to German, since the languages share the same script, we converted German to Fake-German in an analogous method to the Fake-English conversion. Finally, German to Greek ($DeuEll$) experiments were also conducted for completion.

Firstly, we examine the uncontextualized embeddings of characters. To retrieve the representation of a character, we give it as input to the model and extract the embedding layer activations. We also investigate contextualized embeddings by feeding entire words into the model and extracting the activations for a particular character at the 5th layer.

For all the setups, a separate development set was used for evaluation by holding out 30% of examples (ie., the lines in the dataset). After training the respective models, we compute the cosine similarity matrix between the two alphabets.

4 Results

4.1 Similarity Matrix Comparisons

The cosine similarity matrices are shown for our different setups. First, we give as input the characters of the two alphabets separately and extract their first-layer representations. We examined all layer representations, but since the characters are given without context, we decided to go with the first layer, which has been shown to contain context-independent information (Jawahar et al., 2019).

In Figure 1 (a) the cosine similarity matrix for $EngFake_{base}$ is presented. The diagonal shows that the model correctly aligns English with the base Fake-English language. In Figure 1 (b) we see the similarity matrix of $EngFake_{f_1 f_2}$. The strong diagonal indicates that BERT indeed manages to align English with its Fake-English equivalent. The added perturbation ($f_1 f_2$ instead of 'f') is correctly captured by the model as well. The English 'f' has high similarity scores with both f_1 and f_2, with f_1 having a slightly higher score. We also compare against the combined $f_1 f_2$ bigram, by computing its joint representation. We denote this new, combined token simply with 'f' in the matrix.

For $EngDeu$, in Table 2 (a), we see lower performance. The diagonal is still observed, but less prominently than before. The two languages are relatively similar and belong to the same language family, so some similarity on the character level is to be expected. Nevertheless, the similarity matrix shows quite significant noise that obscures the diagonal. Note that even though the diagonal is not the best indicator of performance, the model was overall less efficient than expected.

One of the patterns that emerged is the high compatibility of the matrix's right hand side (German 'w', 'x', 'y' and 'z') with most of the English characters. High similarities can be found between these characters and most English characters across the board. This could be because of the low frequency of these letters in German.[2] Apart from that, there are few other clusters around the matrix, for example 'w', 'x' and 'y' for English and the 'h'-'o' range in German.

When compared against a language further away from English than German, in this case Greek, similarities become even fainter. In Table 2 (b) the similarity matrix for $EngEll$ is presented. We see weaker similarity scores across the board with results mostly random. As we show in Section 4.2, only a few characters were correctly aligned. In English, 'e', 'j', 't' and 'z' have high similarity scores overall, with 'ρ', 'σ' and 'τ' in Greek scoring highly across the board as well.

Finally, in Table 2 (c), $DeuEll$ is shown. There is very little that can be inferred from this matrix, since performance seems to be random.

4.2 Max Alignment Accuracy

Here we take a closer look at the similarity matrices and quantify how well our model aligns characters compared to a ground truth value. Even though ground truth alignment between characters is not always clear (for example, the English 'a' has multiple pronunciations: 'allure', 'ball', 'make', which would arguably map to three distinct characters in Greek), there are some obviously incorrect alignments we can observe (for example, the English 'a' should never be aligned with any of the Greek consonants). Thus, this max alignment method is an adequate indicator of model performance.

In Table 1 we examine alignments for $EngDeu$ and $EngEll$. When choosing a target character, we search in the similarity matrix for the given pair of languages and choose the character with the maximum cosine similarity. In the natural language setups, the model fares fairly badly. In $EngDeu$ the model correctly aligns 11 characters (this is of course a mere simplification, since alignments such

---

[2]https://www.sttmedia.com/characterfrequency-german

18

(a) monastery			(b) test			(c) cardiac		
m	μ	0.63						
o	o	0.50						
n	ν	0.37				c	κ	0.77
a	α	0.49				a	α	0.63
s	η	0.23				r	ρ	0.55
t	τ	0.43	t	τ	0.29	d	ρ	0.19
e	η	0.34	e	ε	0.21	i	ι	0.24
r	ρ	0.31	s	τ	0.13	a	α	0.49
y	ι	0.22	t	σ	0.34	c	κ	0.52

Table 2: Max alignments for 'monastery', 'test' and 'cardiac' as computed by our model.

as 'k' → 'c' could also be considered correct; here we consider only the most basic of alignments: the ones on the diagonal). The situation is worse in $EngEll$, where only 3 characters were correctly aligned. $EngFake$ results are omitted, since the model correctly aligned all characters in all setups.

4.3 Contextualized Embeddings

We also perform a qualitative study on contextualized character-level representations. We choose three pairs of words, feed them separately to the model and extract their 5th-layer representations (averages of all layers were also examined with similar results, as well as other individual layers; the best performing layer was chosen). After computing the contextual representations for all characters in the English word, we align them with the contextual representations of the characters in the Greek word. In Table 2 these alignments are shown.

Results are seemingly random with a strong bias towards the diagonal[3]. The model, thus, has not learned any contextual representations either.

5 Ablation Studies

We also performed some minor experiments with static embeddings and adjustments to the MLM task, as well as a variation to our $EngFake$ setup.

Apart from the BERT embeddings, we also examined static embeddings, to introduce another baseline. Specifically, we used the FastText algorithm (Bojanowski et al., 2017). Data was the same as the previous experiments. Results were seemingly random, with FastText unable to capture any meaningful representations. Cosine similarities were generally very low and alignments incoherent.

For the MLM task, we experimented with different probabilities for token masking. In the origi-

nal BERT paper, the chosen token is replaced by [MASK] 80% of the time, while it remains the same 10% of the time and gets replaced by a random token the rest 10%. We experimented with the following distributions: 80/20/0, 60/20/20 and 50/50/0 with no noticeable change in performance.

We also experimented with overlapping Fake-English mappings in the form 'a' → '100 101 102'. That is, each character is mapped to the tokens corresponding to itself and its next two characters (eg. 'a' is mapped to the tokens for 'a', 'b' and 'c'). This is an extreme case of the $EngFake_{f_1 f_2}$ setup, where each character is mapped to three (instead of 'f' getting mapped to 'f1' and 'f2'). Results deteriorated to random performance.

Finally, we conducted another $EngFake$ experiment. Namely, we restricted the context where we present each character. For the previous experiments each word was split in characters. Now, we further split each word into trigrams. So, instead of 'e x a m p l e' we end up with multiple overlapping trigram entries: 'e x a', 'x a m', 'a m p', etc. The motivation for this experiment is that since the transliteration of a character in a word doesn't need the entire word but instead only its direct neighbors, we should examine a setup with more restricted context[4]. In this case, results deteriorated heavily and are seemingly random with no meaningful representation captured.

6 Conclusion

In our work BERT is shown to be unable to create consistently good cross-lingual spaces on the character level. We train models on English, German, Greek and Fake-English and we compare character-level alignments between them. Cosine similarity matrices between the target and source alphabets were examined and we found that the closer two languages are, the better BERT does in aligning them. Fake-English is the easiest to align with English, whereas German is worse, with Greek trailing far behind. We conclude that BERT is not able to perform adequate character alignment.

Acknowledgments

---

[3]To control for character positions, we also aligned characters without positional embeddings and results got worse.

[4]While this is a more nuanced subject, for the purposes of this ablation experiment we assumed this statement stood.

(a) $EngFake$
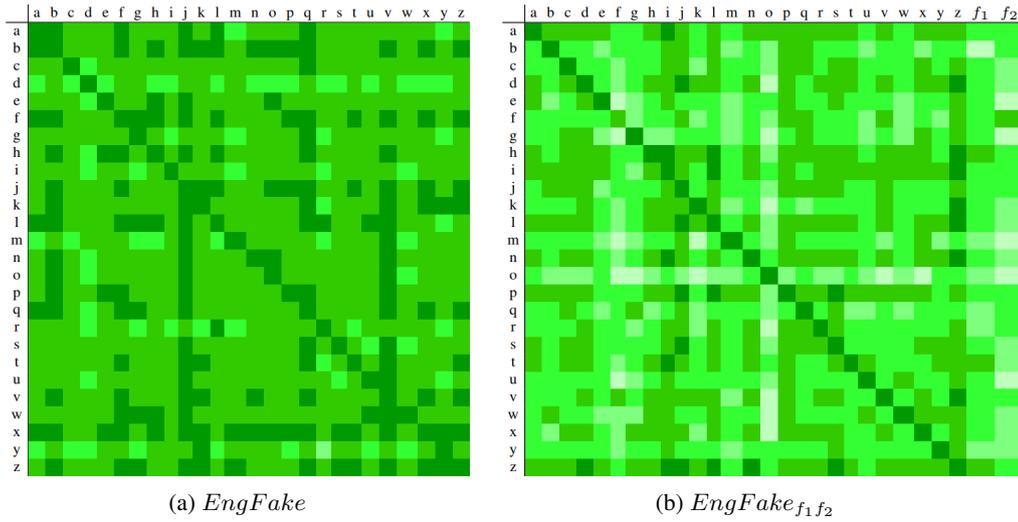
(b) $EngFake_{f_1 f_2}$

Figure 1: Showing heatmaps for (a) $EngFake$ and (b) $EngFake_{f_1 f_2}$. The lighter green cells show lower cosine similarity, the darker green cells show higher cosine similarity.

(a) $EngDeu$
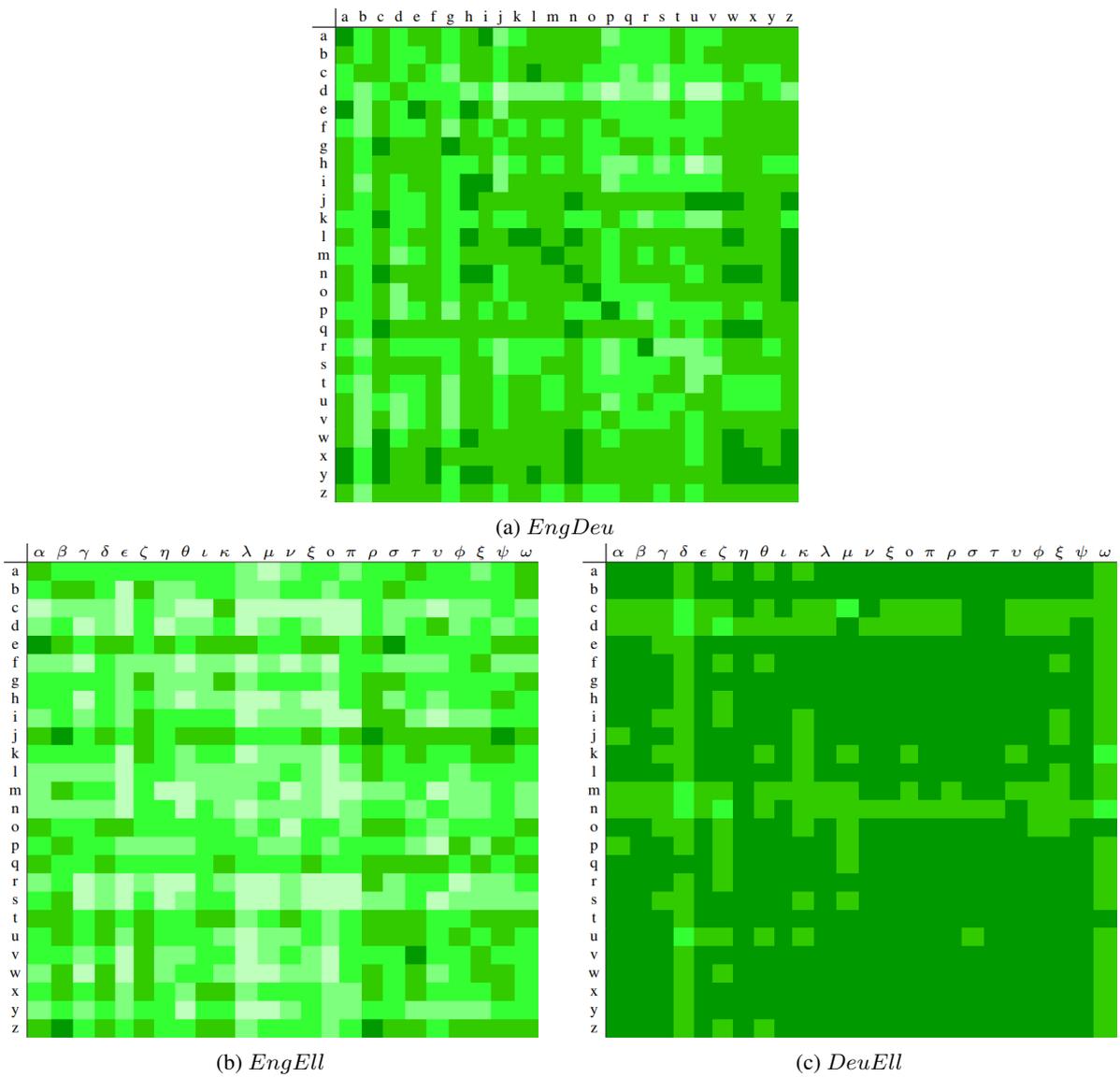
(b) $EngEll$

(c) $DeuEll$

Figure 2: Showing heatmaps for (a) $EngDeu$, (b) $EngEll$ and (c) $DeuEll$. The lighter green cells show lower cosine similarity, the darker green cells show higher cosine similarity.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069. Curran Associates, Inc.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. 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 974–983, Hong Kong, China. 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.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

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

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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 4453–4462, Hong Kong, China. Association for Computational Linguistics.

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Joël Legrand, Michael Auli, and Ronan Collobert. 2016. Neural network-based word alignment through score aggregation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 66–73, Berlin, Germany. Association for Computational Linguistics.

Haizhou Li, A Kumaran, Vladimir Pervouchine, and Min Zhang. 2009. Report of NEWS 2009 machine transliteration shared task. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 1–18, Suntec, Singapore. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

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.

Hassan Sajjad, Helmut Schmid, Alexander Fraser, and Hinrich Schütze. 2017. Statistical models for unsupervised, semi-supervised supervised transliteration mining. *Computational Linguistics*, 43(2):349–375.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

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

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment.