CROSS-LINGUAL ABILITY OF MULTILINGUAL BERT: AN EMPIRICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has exhibited the surprising cross-lingual abilities of multilingual BERT (M-BERT) – surprising since it is trained without any cross-lingual objective and with no aligned data. In this work, we provide a comprehensive study of the contribution of different components in M-BERT to its cross-lingual ability. We study the impact of linguistic properties of the languages, the architecture of the model, and the learning objectives. The experimental study is done in the context of three typologically different languages – Spanish, Hindi, and Russian – and using two conceptually different NLP tasks, textual entailment and named entity recognition. Among our key conclusions is the fact that the lexical overlap between languages plays a negligible role in the cross-lingual success, while the depth of the network is an integral part of it.

1 INTRODUCTION

Embeddings of natural language text via unsupervised learning, coupled with sufficient supervised training data, have been ubiquitous in NLP in recent years and have shown success in a wide range of monolingual NLP tasks, mostly in English. Training models for other languages have been shown more difficult, and recent approaches relied on bilingual embeddings that allowed the transfer of supervision in high resource languages like English to models in lower resource languages; however, inducing these bilingual embeddings required some level of supervision (Upadhyay et al., 2016).

Multilingual BERT[1] (M-BERT), a Transformer-based (Vaswani et al., 2017) language model trained on raw Wikipedia text of 104 languages suggests an entirely different approach. Not only the model is contextual, but its training also requires no supervision – no alignment between the languages is done. Nevertheless, and despite being trained with no explicit cross-lingual objective, M-BERT produces a representation that seems to generalize well across languages for a variety of downstream tasks (Wu & Dredze, 2019).

In this work, we attempt to develop an understanding of the success of M-BERT. We study a range of aspects on a couple of different NLP tasks, in order to identify the key components in the success of the model. Our study is done in the context of only two languages, source (typically English) and target (multiple, quite different languages). By involving only a pair of languages, we can study the performance on a given target language, ensuring that it is influenced only by the cross-lingual transfer from the source language, without having to worry about a third language interfering.

We analyze the two-languages version of M-BERT (B-BERT, from now on) in three orthogonal dimensions: (i) Linguistics properties and similarities of target and source languages; (ii) Network Architecture, and (iii) Input and Learning Objective.

One hypothesis that came up when the people thoughts about the success of M-BERT is due to some level of language similarity. This could be lexical similarity (shared words or word-parts) or structural similarities, or both. We, therefore, investigate the contribution of *word-piece overlap* – the extent to which the same word-pieces appear in both source and target languages – and distinguish it from other similarities, which we call *structural similarity* between the source and target languages. Surprisingly, as we show, B-BERT is cross-lingual even when there is absolutely no word-piece overlap. That is, other aspects of language similarity must be contributing to the cross-lingual capabilities of the model. This is contrary to Pires et al. (2019) hypothesis that M-BERT gains its

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

power from shared word-pieces. Furthermore, we show that the amount of word-piece overlap in B-BERT's training data contributes little to performance improvements.

Our study of the model architecture addresses the importance of (i) The network depth, (ii) the number of attention heads, and (iii) the total number of model parameters in B-BERT. Our results suggest that depth and the total number of parameters of B-BERT are crucial for both monolingual and cross-lingual performance, whereas multi-head attention is not a significant factor – a single attention head B-BERT can already give satisfactory results.

To understand the role of the learning objective and the input representation, we study the effect of (i) the next sentence prediction objective, (ii) the language identifier in the training data, and (iii) the level of tokenization in the input representation (character, word-piece, or word tokenization). Our results indicate that the next sentence prediction objective actually hurts the performance of the model while identifying the language in the input does not affect B-BERT's performance cross-lingually. Our experiments also show that character-level and word-level tokenization of the input results in significantly worse performance than word-piece level tokenization.

Overall, we provide an extensive set of experiments on three source-target language pairs, English–Spanish, English–Russian, and English–Hindi. We chose these target languages since they vary in scripts and typological features. We evaluate the performance of B-BERT on two very different downstream tasks: cross-lingual Named Entity Recognition – a sequence prediction task the requires only local context – and cross-lingual Textual Entailment Dagan et al. (2013) that requires more global representation of the text.

Ours is not the first study of M-BERT. (Wu & Dredze, 2019) and (Pires et al., 2019) identified the cross-lingual success of the model and tried to understand it. The former by considering M-BERT layerwise, relating cross-lingual performance with the amount of shared word-pieces and the latter by considering the model's ability to transfer between languages as a function of word order similarity in languages. However, both works treated M-BERT as a black box and compared M-BERT's performance on different languages. This work, on the other hand, examines how B-BERT performs cross-lingually by probing its components, along multiple aspects.

We also note that some of the architectural conclusions have been observed earlier, if not investigated, in other contexts. Liu et al. (2019); Yang et al. (2019) argued that the next Sentence prediction objective of BERT (the monolingual model) is not very useful; we show that this is the case in the cross-lingual setting. Voita et al. (2019) prunes attention heads for a transformer based machine translation model and argues that most attention heads are not important; in this work, we show that the number of attention heads is not important in the cross-lingual setting.

Our contributions are threefold: (i) we provide the first extensive study of the aspects of the multilingual BERT that give rise to its cross-lingual ability. (ii) We develop a methodology that facilitates the analysis of similarities between languages and their impact on cross-lingual models; we do this by mapping English to a Fake-English language, that is identical in all aspects to English but shares not word-pieces with any target language. Finally, (iii) we develop a set of insights into B-BERT, along linguistics, architectural, and learning dimensions, that would contribute to further understanding and to the development of more advanced cross-lingual neural models.

2 BACKGROUND

2.1 BERT

BERT (Devlin et al., 2019) is Transformer (Vaswani et al., 2017) based pre-training language representation model that has been widely used in the field of Natural Language Processing. BERT is trained with Masked Language Modelling (MLM) (Taylor, 1953) and Next Sentence Prediction (NSP) objectives. Input to BERT is a pair of sentences[2] A and B, such that half of the time B comes after A in the original text and the rest of the time B is a randomly sampled sentence. Some tokens from the input are randomly masked, and the MLM objective is to predict the masked tokens. The NSP objective is to predict whether B is the actual next sentence of A or not. (Devlin et al., 2019) argues that MLM enables a deep representation from both directions, and NSP helps understand the relationship between two sentences and can be beneficial to representations.

[2]In the implementation, the input is a pair of segments, which can contain multiple sentences

BERT follows two-steps 1. Pre-training and, 2. Fine-tuning. BERT is pre-trained using the above mentioned MLM and NSP objective on BooksCorpus and English Wikipedia text, and for any supervised downstream task, BERT is initialized with the pre-trained weights and fine-tuned using the labeled data. BERT uses wordpiece tokenization (Wu et al., 2016), which creates wordpiece vocabulary in a data-driven approach.

2.2 MULTILINGUAL BERT

Multilingual BERT is pre-trained in the same way as monolingual BERT except using Wikipedia text from the top 104 languages. To account for the differences in the size of Wikipedia, some languages are sub-sampled, and some are super-sampled using exponential smoothing Devlin et al. (2018). It's worth mentioning that there are no cross-lingual objectives specifically designed nor any cross-lingual data, e.g. parallel corpus, used.

3 WHY MULTILINGUAL BERT WORKS

In this section, we analyze the reason for the cross-lingual ability of multilingual BERT (actually B-BERT) in three dimensions. (i) Linguistics (ii) Architecture and (iii) Input and Learning Objective. Languages share similarities with each other. For example, English and Spanish have words that look seemingly the same; English and Russian both have a Subject-Verb-Object(SVO) order[3]; English and Hindi, despite in different scripts, use the same Arabic numerals[4]. The similarity between languages can be a reason of M-BERT's cross-lingual ability. In this linguistics point of view, we study the contribution of word-piece overlap — the similarity of languages arising from the same characters/words used across languages as well as code-switching data — and structure similarity, the part of linguistic similarity that is not explained by word-piece overlap, and does not rely on the script of the language

We hypothesize that the cross-lingual effectiveness of B-BERT comes from the architecture of BERT itself being able to extract good semantic and structural features. We study the depth, number of attention heads, and the total number of parameters of the transformer model to explore the influence of each part to the cross-lingual ability.

Finally, we study the effect of learning objectives and input. The Next Sentence Prediction objective is shown to be unnecessary in monolingual settings, and we try to analyze its effect in the cross-lingual setting. B-BERT follows BERT and uses a word-piece vocabulary. Word-pieces can be seen as a tradeoff between characters and words. We compare these three ways of tokenizing the input on how they affect cross-lingual transferring.

3.1 DATASETS AND EXPERIMENTAL SETUP

In this work, we conduct all our experiments on two conceptually different downstream tasks – cross-lingual Textual Entailment (TE) and cross-lingual Named Entity Recognition (NER). TE measures natural language understanding (NLU) at a sentence and sentence pair level, whereas NER measures NLU at a token level. We use the Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018) dataset to evaluate cross-lingual TE performance and LORELEI dataset (Strassel & Tracey, 2016) for Cross-Lingual NER.

3.1.1 CROSS-LINGUAL NATURAL LANGUAGE INFERENCE (XNLI)

XNLI is a standard cross-lingual textual entailment dataset that extends MultiNLI (Williams et al., 2018) dataset by creating a new dev and test set and manually translating into 14 different languages. Each input consists of a premise and hypothesis pair, and the task is to classify the relationship between premise and hypothesis into one of the three labels: entailment, contradiction, and neutral. While training, both premise, and hypotheses are in English, and while testing, both are in the target language. XNLI uses the same set of premises and hypotheses for all the language, making the comparison across languages possible.

3.1.2 CROSS-LINGUAL NAMED ENTITY RECOGNITION (NER)

Named Entity Recognition is the task of identifying and labeling text spans as named entities, such as people's names and locations. The NER dataset (Strassel & Tracey, 2016) we use consists of news and social media text labeled by native speakers following the same guideline in several languages,

[3]Although Russian uses flexible word ordering, SVO ordering is predominant

[4]Hindi has its own numerals but used less frequently

including English, Hindi, Spanish, and Russian. We subsample 80%, 10%, 10% of English NER data as training, development, and testing. We use the whole dataset of Hindi, Spanish, and Russian for testing purposes. The vocabulary size is fixed at 60000 and is estimated through the unigram language model in the SentencePiece library (Kudo, 2018).

3.1.3 NOTATION AND EXPERIMENTAL SETUP

We denote B-BERT trained on language A and B as A-B, e.g., B-BERT trained on English (en) and Hindi (hi) as en-hi, similarly for Spanish (es) and Russian (ru). For pretraining, we subsample en, es, and ru Wikipedia to 1GB and use the entire Wikipedia for Hindi. Unless otherwise specified, for B-BERT training, we use a batch size of 32, the learning rate of 0.0001, and 2M training steps.

For XNLI, we use the same finetuning approach as BERT uses in English and report accuracy. For NER, we extract BERT representations as features and finetune a Bi-LSTM CRF model and report entity span F_1 score averaged from 5 runs with its standard deviation.

3.2 LINGUISTICS

Pires et al. (2019) hypothesizes that the cross-lingual ability of M-BERT arises because of the shared word-pieces between source and target languages. However, our experiments show that B-BERT is cross-lingual even when there is no word-piece overlap. Further, (Wu & Dredze, 2019) hypothesizes that for cross-lingual transfer learning source language should be selected such that it shares more word-pieces with the target language. However, our experiment suggests that structural similarity is much more important. Motivated by the above two hypotheses, in this section, we study the contribution of word-piece overlap and structural similarity for the cross-lingual ability of B-BERT.

3.2.1 WORD-PIECE OVERLAP

M-BERT model is trained using Wikipedia text from 104 languages, and the texts from different languages share some common wordpiece vocabulary (like numbers, links, etc.. including actual words, if they have the same script), we refer to this as word-piece overlap. The previous work (Pires et al., 2019) hypothesizes that M-BERT generalizes across languages because these shared word-pieces have to be mapped to shared space forcing the other co-occurring word-pieces to be mapped to the same shared space.

In this section, we perform experiments to compare cross-lingual performance with and without word-piece overlap. We construct a new corpus – *Fake-English (enfake)*, by shifting the Unicode of each character in English Wikipedia text by a large constant so that there is strictly no character overlap with any other Wikipedia text. In this work, we consider Fake-English as a different language.

			XNLI		**NER**
B-BERT	**Train**	**Test**	**Accuracy**	**Wordpiece Contribution**	**Span F1-Score**
en-es	en	es	72.3	1.4	61.9 ($\pm$0.8)
enfake-es	enfake		**70.9**		**62.6 ($\pm$1.6)**
en-hi	en	hi	60.1	0.5	61.6 ($\pm$0.7)
enfake-hi	enfake		**59.6**		**62.9 ($\pm$0.7)**
en-ru	en	ru	66.4	0.7	57.1* ($\pm$0.9)
enfake-ru	enfake		**65.7**		**54.2 ($\pm$0.7)**
en-enfake	enfake	enfake	78.0	0.5	78.9*($\pm$0.7)
en-enfake	enfake	en	**77.5**		76.6($\pm$0.8)

Table 1: **The Effect of Word-Piece Overlap and of Structural Similarity** For different pairs of B-BERT languages, and for two tasks (XNLI and NER), we show the contribution of word-pieces to the success of the model. In every two consecutive rows, we show results for a pair (e.g., English-Spanish) and then for the corresponding pair after mapping English to a disjoint set of word-pieces. The gap between the performance in each group of two rows indicates the loss due to completely eliminating the word-piece contribution. We add an asterisk to the number for NER when the results are statistically significant at the 0.05 level.

We measure the contribution of word-piece overlap as the drop in performance when the word-piece overlap is removed. From Table 1, we can see B-BERT *is cross-lingual even when there is no word-piece overlap*. We can also see that the contribution of word-piece overlap is very small, which is quite surprising and contradictory to the hypothesis by (Pires et al., 2019; Wu & Dredze, 2019).

3.2.2 STRUCTURAL SIMILARITY

We define the structure of a language as every property of an individual language that is invariant to the script of the language, e.g., morphology, word-ordering, word frequency, word-pair frequency are all part of the structure of a language. Note that English and Fake-English don't share any vocabulary/characters, but they have exactly the same structure. From Table 1, we can see that BERT transfers very well from Fake-English to English. Also note that, despite not sharing any vocabulary, Fake-English transfers to Spanish, Hindi, Russian almost as well as English. On XNLI, where the scores between languages can be compared, the cross-lingual transferability from Fake-English to Spanish is much better than from Fake-English to Hindi/Russian. Since they do not share any word-pieces, this better transferability comes from the structure being closer between Spanish and Fake-English. These results suggest that we should shed more light on studying the structural similarity between languages. In this study, we don't further dissect the structure of language as currently, the definition of "Structure of a Language" is fuzzy. Despite its amorphous definition, our experiment clearly shows that structural similarity is crucial for cross-lingual transfer.

3.3 ARCHITECTURE

From Section 3.2, we observe that B-BERT recognizes the language structure effectively, We envisage that BERT potentially gains the ability to recognize language structure because of its architecture. In this section, we study the contribution of different components of B-BERT architecture namely (i) depth, (ii) multi-head attention and, (iii) the total number of parameters. The motivation is to understand which components are crucial for its cross-lingual ability.

We perform all our cross-lingual experiments on the XNLI dataset with Fake-English as the source and Russian as the target language; we measure cross-lingual ability by the difference between the performance of Fake-English and Russian (lesser the difference better the cross-lingual ability).

3.3.1 DEPTH

We presume the ability of B-BERT to extract good semantic and structural features is a crucial reason for its cross-lingual effectiveness, and the deepness of B-BERT helps it extract good language features. In this section, we study the effect of depth on both the monolingual and cross-lingual performance of B-BERT. We fix the number of attention heads and change the size of hidden units and intermediate units such that the total number of parameters are almost the same (size of intermediate units is always $4\times$ size of hidden units).

			XNLI		
Parameters (in Millions)	**Depth**	**Multi-head Attention**	**Fake-English**	**Russian**	Δ
138.69	1	12	66.6	45.0	21.6
136.32	2	12	73.7	55.7	18.0
136.20	4	12	76.9	59.0	17.9
138.86	6	12	78.3	63.1	15.2
134.00	8	12	78.6	62.3	16.3
136.10	18	12	79.1	66.0	13.1
139.33	24	12	78.9	67.6	11.3
132.78	12	12	79.0	65.7	13.3

Table 2: **The Effect of Depth of B-BERT Architecture:** We use Fake-English and Russian B-BERT and study the effect of depth of B-BERT on the performance of Fake-English and the Russian language on XNLI data. We vary the depth and fix both the number of attention heads and the number of parameters – the size of hidden and intermediate units are changed so that the total number of parameters remains almost the same. We train only on Fake-English and test on both Fake-English and Russian and report their test accuracy. The difference between the performance on Fake-English and Russian(Δ) is our measure of cross-lingual ability (lesser the difference, better the cross-lingual ability).

From Table 2, we can see that a deeper model not only perform better on English but are also better cross-lingual(Δ). We can also see a strong correlation between performance on English and cross-lingual ability (Δ), which further supports our assumption that the ability to extract good semantic and structural features is a crucial reason for its cross-lingual effectiveness.

3.3.2 MULTI-HEAD ATTENTION

In this section, we study the effect of multi-head attention on the cross-lingual ability of B-BERT. We fix the depth and the total number of parameters – which is a function of depth and size of

hidden and intermediate and study the performance for the different number of attention heads. From Table 3, we can see that the number of attention heads doesn't have a significant effect on cross-lingual ability(Δ) – B-BERT is satisfactorily cross-lingual even with a single attention head, which is in agreement with the recent study on monolingual BERT (Voita et al., 2019; Clark et al., 2019).

			XNLI		
Parameters (in Millions)	**Depth**	**Multi-head Attention**	**Fake-English**	**Russian**	Δ
132.78	12	1	77.4	63.2	14.2
132.78	12	2	78.3	62.8	15.5
132.78	12	3	79.5	65.3	14.2
132.78	12	6	78.9	66.7	12.2
132.78	12	16	77.9	64.9	13.0
132.78	12	24	77.9	63.9	14.0
132.78	12	12	79.0	65.7	13.3

Table 3: **The Effect of Multi-head Attention:** We study the effect of the number of attention heads of B-BERT on the performance of Fake-English and Russian language on XNLI data. We fix both the number of depth and number of parameters of B-BERT and vary the number of attention heads. The difference between the performance on Fake-English and Russian(Δ) is our measure of cross-lingual ability.

3.3.3 Total Number of Parameters

Similar to the depth, we also anticipate that a large number of parameters could potentially help B-BERT extract good semantic and structural features. We study the effect of the total number of parameters on cross-lingual performance by fixing the number of attention heads and depth; we change the number of parameters by changing the size of hidden and intermediate units (size of intermediate units is always $4\times$ size of hidden units). From Table 4, we can see that the total number of parameters is not as significant as depth; however, below a threshold, the number of parameters seems significant, which suggests that B-BERT requires a certain minimum number parameters to extract good semantic and structural feature.

			XNLI		
Parameters (in Millions)	**Depth**	**Multi-head Attention**	**Fake-English**	**Russian**	Δ
7.87	3	3	68.5	43.2	25.3
12.19	3	3	70.1	44.1	26.0
16.78	3	3	70.8	50.4	20.4
8.40	6	6	70.2	49.7	20.5
13.37	6	6	72.4	56.2	16.2
18.87	6	6	73.3	54.4	18.9
29.65	12	12	76.6	61.4	15.2
44.89	12	12	78.2	64.0	14.2
89.03	12	12	78.6	64.1	14.5
283.11	12	12	79.6	65.4	14.2
132.78	12	12	79.0	65.7	13.3

Table 4: **The Effect of Total Number of Parameters:** We study the effect of the total number of Parameters of B-BERT on the performance of Fake-English and Russian language on XNLI data. We fix both the number of depth and number of attention heads of B-BERT and vary the total number of parameters by changing the size of hidden and intermediate units. The difference between the performance on Fake-English and Russian(Δ) is our measure of cross-lingual ability.

3.4 Input and Learning Objective

In this section, we study the effect of input representation and learning objectives on the cross-lingual ability of B-BERT. BERT is a Transformer model trained with MLM and NSP objectives. XLM (Lample & Conneau, 2019) shows that the Transformer model trained with Causal Language Modeling (CLM) objective is also cross-lingual; however, it also observes that pre-training with MLM objective consistently outperforms the one with CLM. In this work, we don't study further the effect of MLM objective. Recent works (Lample & Conneau, 2019; Joshi et al., 2019; Liu et al., 2019) show that the NSP objective hurts the performance of several monolingual tasks; in this work, we verify if the NSP objective helps or hurts cross-lingual performance.

M-BERT (Devlin et al., 2018) states that they intentionally do not use any marker to identify language so that cross-lingual transfer works, however, our experiments suggest that adding a language identity marker to the input doesn't hurt the cross-lingual performance of BERT. We are also interested in studying the effect of characters and words vocabulary instead of word-pieces. Characters provide handling unseen words better than words, words carry more semantic and syntactic information inside it, and word-pieces is more of a middle ground of these two.

3.4.1 NEXT SENTENCE PREDICTION (NSP)

The input to the BERT is a pair of sentences separated by a special token such that half the time the second sentence is the next and rest half the time, it is a random sentence. The NSP objective of BERT (B-BERT) is to predict whether the second sentence comes after the first one in the original text. We study the effect of NSP objective by comparing the performance of B-BERT pre-trained with and without this objective. From Table 5, we can see that the NSP objective hurts the cross-lingual performance even more than monolingual performance.

			XNLI		**NER**	
B-BERT	**Train**	**Test**	**NSP**	**No-NSP**	**NSP**	**No-NSP**
enfake-es	enfake	enfake	78.5	78.7	80.3 (±0.6)	80.7 (±1.4)
		es	70.9	72.7	62.6 (±1.6)	64.6 (±1.4)
enfake-hi	enfake	enfake	79.3	80.1	81.4 (±0.9)	80.0 (±1.1)
		hi	59.6	60.7	62.9 (±0.7)	62.4 (±1.4)
enfake-ru	enfake	enfake	79.0	79.0	80.2 (±0.7)	80.3 (±0.8)
		ru	65.7	66.7	54.2 (±0.7)	55.7 (±0.3)

Table 5: **Effect of Next Sentence Prediction Objective:** We study the effect of NSP objective on XNLI and NER. Column NSP and No-NSP show the performance (accuracy for XNLI and average (stdev) F1-score for NER) when B-BERT is trained with and without NSP objective respectively. The difference between the NSP and No-NSP shows that NSP objective hurts performance.

3.4.2 LANGUAGE IDENTITY MARKER

In this work, we argue that B-BERT is cross-lingual because of its ability to recognize language structure and semantics, and hence we presume adding a language identity marker doesn't affect its cross-lingual ability. Even if we don't add a language identity marker, BERT learns language identity (Wu & Dredze, 2019). To incorporate language identity in the input, we add a different end of string tokens([SEP]) for different languages (i.e., our input format is [CLS] SENT1 [SEP-M] SENT2 [SEP-N], where M and N are languages corresponding to SENT1 and SENT2 respectively). From Table 6, we can observe that adding language identity marker doesn't affect cross-lingual performance.

			XNLI		**NER**	
B-BERT	**Train**	**Test**	**No Lang-id**	**With Lang-id**	**No Lang-id**	**With Lang-id**
enfake-es	enfake	enfake	78.5	78.4	80.3 (±0.6)	81.7 (±1.1)
		es	70.9	72.2	62.6(±1.6)	62.2 (±0.4)
enfake-hi	enfake	enfake	79.3	79.0	81.4 (±0.9)	80.7 (±1.6)
		hi	59.6	59.6	62.9 (±0.7)	61.0 (±0.7)
enfake-ru	enfake	enfake	79.0	78.4	80.2 (±0.7)	79.1 (±1.8)
		ru	65.7	65.3	54.2 (±0.7)	55.7 (±0.6)

Table 6: **Effect of Language Identity Marker in the Input:** We study the effect of adding a language identifier in the input data. We use different end of string([SEP]) tokens for different languages serving as language identity marker. Column "With Lang-id" and "No Lang-id" show the performance when B-BERT is trained with and without language identity marker in the input.

3.4.3 CHARACTER VS. WORD-PIECE VS. WORD

In this section, we compare the performance of B-BERT with character, word-piece, and word tokenized input. For character B-BERT, we use all the characters as vocabulary, and for word B-BERT, we use the most frequent 100000 words as vocabulary. From Table 7, we can see that both monolingual and cross-lingual performance of B-BERT with word-piece tokenized input is

better than the character as well as word tokenized input. We believe that this is because word-pieces carry much more information than characters, and word-pieces address unseen words better than words.

			XNLI			NER		
B-BERT	**Train**	**Test**	**Char**	**WordPiece**	**Word**	**Char**	**WordPiece**	**Word**
enfake-es	enfake	enfake	73.7	80.0	80.3	78.8 ($\pm$1.3)	80.3 ($\pm$1.5)	74.9 ($\pm$ 2.2)
		es	66.6	74.9	74.4	62.0 ($\pm$0.8)	64.8 ($\pm$0.9)	57.5 ($\pm$0.4)
enfake-hi	enfake	enfake	73.9	80.3	80.0	79.6 ($\pm$0.9)	79.7 ($\pm$1.1)	75.0 ($\pm$ 1.9)
		hi	53.8	61.7	60.3	53.1 ($\pm$0.4)	58.8 ($\pm$1.2)	56.6 ($\pm$0.8)
enfake-ru	enfake	enfake	74.2	80.7	79.2	77.2 ($\pm$1.1)	80.8 ($\pm$1.3)	73.8 ($\pm$ 0.9)
		ru	61.4	68.1	65.0	52.1 ($\pm$0.5)	56.5 ($\pm$0.3)	46.4 ($\pm$1.3)

Table 7: **Effect of Character vs Word-Piece vs Word tokenization**. We compare the performance of B-BERT with different tokenized input on XNLI and NER data. Column *Char, WordPiece, Word* reports the performance of B-BERT with character, wordpiece and work tokenized input respectively. We use 2k batch size and 500k epochs.

4 Discussion and Future work

This paper provides a systematic empirical study addressing the cross-lingual ability of B-BERT. The analysis presented here covers three dimensions: (1) Linguistics properties and similarities of the source and target languages, (2) Neural Architecture, and (3) Input representation and Learning Objective.

In order to gauge the language similarity aspect needed to make B-BERT successful, we created a new language – Fake-English – and this allows us to study the effect of word-piece overlap while maintaining all other properties of the source language. Our experiments reveal some interesting and surprising results like the fact that word-piece overlap on the one hand, and multi-head attention on the other, are both not significant, whereas structural similarity and the depth of B-BERT are crucial for its cross-lingual ability.

While, in order to better control interference among languages, we studied the cross-lingual ability of B-BERT instead of those of M-BERT, it would be interesting now to extend this study, allowing for more interactions among languages. We leave it to future work to study these interactions. In particular, one important question is to understand the extent to which adding to M-BERT languages that are *related* to the target language, helps the model's cross-lingual ability.

We introduced the term *Structural Similarity*, despite its obscure definition, and show its significance in cross-lingual ability. Another interesting future work could be to develop a better definition and, consequently, a finer set of experiments, to better understand the *Structural similarity* and study its individual components.

Finally, we note an interesting observation made in Table 8. We observe a drastic drop in the entailment performance of B-BERT when the premise and hypothesis are in different languages. (This data was created using XNLI when in the original form the languages contain same premise and hypothesis pair). One of the possible explanations could be that BERT is learning to make textual entailment decisions by matching words or phrases in the premise to those in the hypothesis. This question, too, is left as a future direction.

		Premise Language – Hypothesis Language (XNLI)			
B-BERT	**Target**	**enfake-target**	**target-enfake**	**enfake-enfake**	**target-target**
enfake-es	es	**57.9**	**61.1**	78.5	70.9
enfake-hi	hi	**45.7**	**55.6**	79.3	59.6
enfake-ru	ru	**51.1**	**57.9**	79.0	65.7

Table 8: **Premise and Hypothesis in different language:** Using XNLI test set, we construct textual entailment data with premise and hypothesis in different languages. The column A-B (e.g. enfake-target) refers to test data with premise in language A (enfake) and hypothesis in language B (target). We always train on Fake-English and report test accuracy.

REFERENCES

Anonymous. Reweighted proximal pruning for large-scale language representation. In *Submitted to International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gBOxSFwr. under review.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. In *BlackBoxNLP@ACL*, 2019. URL https://nlp.stanford.edu/pubs/clark2019what.pdf.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://www.aclweb.org/anthology/D18-1269.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzoto. Recognizing textual entailment: Models and applications. 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Multilingual bert - r, 2018. URL https://github.com/google-research/bert/blob/master/multilingual.md.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 66–75, 2018. doi: 10.18653/v1/P18-1007. URL https://www.aclweb.org/anthology/P18-1007/.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

Stephanie Strassel and Jennifer Tracey. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 3273–3280, 2016.

Wilson L Taylor. cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30 (4):415–433, 1953.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5797–5808, 2019. URL https://www.aclweb.org/anthology/P19-1580/.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://www.aclweb.org/anthology/N18-1101.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*, 2019.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

A Appendix

A.1 Further discussions on Structural Similarity

In the main text, we defined structural similarity as all the properties of a language that is invariant to the script of the language, like morphology, word-ordering, word-frequency, etc.. Here, we analyze 2 sub-components of structural similarity – word-ordering similarity and word-frequency (Unigram frequency) similarity to understand the concept of structural similarity better.

A.1.1 Word-Ordering Similarity

Words are ordered differently between languages. For example, English has a Subject-Verb-Object order, while Hindi has a Subject-Object-Verb order. We analyze whether similarity in how words are ordered affects learning cross-lingual transferability. We study the effect of word-ordering similarity – one component of structural similarity – by destroying the word-ordering structure by shuffling some percentage of random words in sentences during pretraining. We shuffle both source (fake-English) and target language (permuting any one of them would also be sufficient). This way, the similarity of word-ordering is hidden from B-BERT. We model how much we permute each sentence, for example, when the sentence is 100% permuted, each sentence can be treated as a Bag of Words. For each word, the words that appear in its context (other words in the sentence) is not changed.

Note that during fine-tuning on source language (and testing on target language), we do not permute, as the cross-lingual ability is gained only during the pretraining process.

From Table 9, we can see that the performance drops significantly when we destroy word-order similarity. However, the cross-lingual performance is still quite good, which indicates that there are other components of structural similarity, which could contribute to the cross-lingual ability of B-BERT.

B-BERT	Train	Test	XNLI (acc)	NER (Span F1-Score)
enfake-es	enfake	es	70.9	62.6 (±1.6)
enfake-es-permute-0.25	enfake	es	68.9	42.3 (± 0.6)
enfake-es-permute-0.50	enfake	es	65.5	41.3 (±3.8)
enfake-es-permute-1.00	enfake	es	62.5	39.4 (±1.5)
enfake-hi	enfake	hi	59.6	62.9 (±0.7)
enfake-hi-permute-0.25	enfake	hi	51.4	24.8 (±2.0)
enfake-hi-permute-0.50	enfake	hi	48.3	8.7 (±1.9)
enfake-hi-permute-1.00	enfake	hi	43.1	2.9 (±1.9)
enfake-ru	enfake	ru	65.7	54.2 (±0.7)
enfake-ru-permute-0.25	enfake	ru	63.6	23.6 (±2.3)
enfake-ru-permute-0.50	enfake	ru	59.7	16.3 (±0.4)
enfake-ru-permute-1.00	enfake	ru	53.6	16.3 (±0.9)

Table 9: **Contribution of Word-Ordering similarity:** We study the importance of word-order similarity by analysing the performance of XNLI and NER when some percent of word-order similarity is destroyed. For a certain percent p, we randomly shuffle $p * L$ words in the sentence, where L is the number of total number of words in that sentence. For example, enfake-es-permute-0.25 indicates that in each sentence a random 25% of words are shuffled and enfake-es-permute-1.00 indicates every word in the sentence is randomly shuffled, similarly for others. We can see that word-order similarity is quite important, however there must be other components of structural similarity that could contribute for the cross-lingual ability, as the the performance of enfake-es-permute-1.00 is still pretty good.

A.1.2 Word-frequency Similarity

Here, we study whether only word frequency allows for good cross-lingual representations – no, it is not much useful. We collect the frequency of words in the target language and generate a new monolingual corpus by sampling words based on the frequency, i.e., each sentence is a set of random words sampled from the same unigram frequency as the original target language. The only information about target language BERT has is its unigram frequency and sub-word level information. We train B-BERT using Fake-English and this newly generated target corpus. From Table 10, we can see that the performance is very low, although non-trivial (random performance is 33.33%). Therefore, unigram frequencies alone don't contain enough information for cross-lingual learning (bi-gram or tri-gram frequencies might be useful).

B-BERT	Train	Test	XNLI (acc)	NER (Span F1-Score)
enfake-es	enfake	es	70.9	62.6 (±1.6)
enfake-es-freq	enfake	es	35.4	15.0(± 1.1)
enfake-hi	enfake	hi	59.6	62.9 (±0.7)
enfake-hi-freq	enfake	hi	36.3	1.2 (±0.9)
enfake-ru	enfake	ru	65.7	54.2 (±0.7)
enfake-ru-freq	enfake	ru	34.4	5.7 (±0.5)

Table 10: **Cross-lingual ability from only Unigram frequency:** We study if only unigram frequency is useful for cross-lingual transfer. enfake-es-freq indicates B-BERT trained with Fake-English and the new created corpus, where each sentence is a set of random words sampled from the same unigram distribution as es. The results show that only unigram frequency is not enough for a reasonable cross-lingual performance.

A.2 Further discussions on Architecture

A.2.1 Multilingual BERT

All the experiments in the main text are conducted on Bilingual BERT. However, the results hold even for the multilingual case; to further illustrate this, we experiment on four language BERT (en, es, hi, ru). From table 11, we can see that the performance on XNLI is comparable even with just 15% of parameters, and just 1, 3 attention heads when the depth is good enough, which is in agreement with our observations in the main text.

Also, Currently, there is a lot of interest in reducing the size of BERT (Anonymous, 2020), using various pruning techniques. Our analysis shows that we can get comparable performance by

maintaining (or increasing) depth. Indeed, by changing the number of parameters from 132.78M to 20.05M (approx 85% less) by reducing only the hidden layer sizes, the English performance drops only from 79.0 to 75.0 (about 4%), which is comparable to the drop in MNLI performance of (Anonymous, 2020), where the performance drops by -4.7% for 88.4% reduction of parameters (English XNLI is similar to MNLI). We believe that these pruning techniques could be combined with the insights from the paper to get much better results.

			XNLI			
Parameters (in Millions)	**Depth**	**Multi-head Attention**	**en**	**es**	**hi**	**ru**
132.78 (100%)	12	12	79.0	70.0	50.7	65.2
20.05 (15.1%)	16	1	74.1	63.2	49.6	58.9
20.05 (15.1%)	16	3	75.0	65.3	51.1	60.3
24.20 (18.23%)	24	1	75.0	64.5	48.3	58.6
24.20 (18.23%)	24	3	75.7	65.5	48.9	60.2
30.43 (22.92%)	36	1	75.2	66.5	47.1	60.3
30.43 (22.92%)	36	3	76.0	64.8	47.8	59.8

Table 11: **Significantly smaller multilingual BERT:** We show that the insights derived from bilingual BERT is valid even in the case of multilingual BERT (4 language BERT). Further, we also show that with enough depth, we only need a fewer number of parameters and attention heads to get comparable results.

A.2.2 Number of parameters

Here we show more results on the effect of the number of parameters to get more insights on the threshold on the number of parameters. The experimental setting was similar to that of table 4.

From table 12, we can notice a drastic change in the performance of Russian when the number of parameters is decreased from 11.83M to 7.23M, so we can consider this range to be the required minimum number of parameters, at least for the 12 layer and 12 attention heads situation.

			XNLI		
Parameters (in Millions)	**Depth**	**Multi-head Attention**	**Fake-English**	**Russian**	Δ
4.23	12	12	65.5	46.6	18.9
7.23	**12**	**12**	68.5	**49.9**	18.6
11.83	**12**	**12**	71.3	**56.3**	15.0
15.71	12	12	72.9	58.2	14.7
23.03	12	12	74.8	60.9	13.9
29.65	12	12	76.6	61.4	15.2
44.89	12	12	78.2	64.0	14.2
89.03	12	12	78.6	64.1	14.5
283.11	12	12	79.6	65.4	14.2
132.78	12	12	79.0	65.7	13.3

Table 12: **The Effect of Total Number of Parameters:** We study the effect of the total number of Parameters of B-BERT on the performance of Fake-English and Russian language on XNLI data. We fix both the number of depth and number of attention heads of B-BERT and vary the total number of parameters by changing the size of hidden and intermediate units. The difference between the performance on Fake-English and Russian(Δ) is our measure of cross-lingual ability. We can see that there is a drastic change in performance when we reduce the number of parameters from 11.83M to 7.23M, so this is threshold for kind of a number of parameters.