Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT

Shijie Wu and **Mark Dredze**
Department of Computer Science
Johns Hopkins University
shijie.wu@jhu.edu, mdredze@cs.jhu.edu

Abstract

Pretrained contextual representation models (Peters et al., 2018; Devlin et al., 2019) have pushed forward the state-of-the-art on many NLP tasks. A new release of BERT (Devlin, 2018) includes a model simultaneously pretrained on 104 languages with impressive performance for zero-shot cross-lingual transfer on a natural language inference task. This paper explores the broader cross-lingual potential of mBERT (multilingual) as a zero-shot language transfer model on 5 NLP tasks covering a total of 39 languages from various language families: NLI, document classification, NER, POS tagging, and dependency parsing. We compare mBERT with the best-published methods for zero-shot cross-lingual transfer and find mBERT competitive on each task. Additionally, we investigate the most effective strategy for utilizing mBERT in this manner, determine to what extent mBERT generalizes away from language-specific features, and measure factors that influence cross-lingual transfer.

1 Introduction

Pretrained language representations with self-supervised objectives have become standard in a variety of NLP tasks (Peters et al., 2018; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2019), including sentence-level classification (Wang et al., 2018), sequence tagging (e.g. NER) (Tjong Kim Sang and De Meulder, 2003) and SQuAD question answering (Rajpurkar et al., 2016). Self-supervised objectives include language modeling, the cloze task (Taylor, 1953) and next sentence classification. These objectives continue key ideas in word embedding objectives like CBOW and skip-gram (Mikolov et al., 2013a).

At the same time, cross-lingual embedding models have reduced the amount of cross-lingual supervision required to produce reasonable models; Conneau et al. (2017); Artetxe et al. (2018) use identical strings between languages as a pseudo bilingual dictionary to learn a mapping between monolingual-trained embeddings. Can jointly training contextual embedding models over multiple languages without explicit mappings produce an effective cross-lingual representation? Surprisingly, the answer is (partially) yes. BERT, a recently introduced pretrained model (Devlin et al., 2019), offers a multilingual model (mBERT) pretrained on concatenated Wikipedia data for 104 languages *without any cross-lingual alignment* (Devlin, 2018). mBERT does surprisingly well compared to cross-lingual word embeddings on zero-shot cross-lingual transfer in XNLI (Conneau et al., 2018), a natural language inference dataset. **Zero-shot cross-lingual transfer**, also known as single-source transfer, refers *trains and selects* a model in a source language, often a high resource language, then transfers directly to a target language.

While XNLI results are promising, the question remains: does mBERT learn a cross-lingual space that supports zero-shot transfer? We evaluate mBERT as a zero-shot cross-lingual transfer model on five different NLP tasks: natural language inference, document classification, named entity recognition, part-of-speech tagging, and dependency parsing. We show that it achieves competitive or even state-of-the-art performance with the recommended fine-tune all parameters scheme (Devlin et al., 2019). Additionally, we explore different fine-tuning and feature extraction schemes and demonstrate that with parameter freezing, we further outperform the suggested fine-tune all approach. Furthermore, we explore the extent to which mBERT generalizes away from a specific language by measuring accuracy on language ID

Code is available at https://github.com/shijie-wu/crosslingual-nlp using each layer of mBERT. Finally, we show how subword tokenization influences transfer by measuring subword overlap between languages.

2 Background

(Zero-shot) Cross-lingual Transfer Cross-lingual transfer learning is a type of transductive transfer learning with different source and target domain (Pan and Yang, 2010). A cross-lingual representation space is assumed to perform the cross-lingual transfer. Before the widespread use of cross-lingual word embeddings, task-specific models assumed coarse-grain representation like part-of-speech tags, in support of a delexicalized parser (Zeman and Resnik, 2008). More recently cross-lingual word embeddings have been used in conjunction with task-specific neural architectures for tasks like named entity recognition (Xie et al., 2018), part-of-speech tagging (Kim et al., 2017) and dependency parsing (Ahmad et al., 2019).

Cross-lingual Word Embeddings. The quality of the cross-lingual space is essential for zero-shot cross-lingual transfer. Ruder et al. (2017) surveys methods for learning cross-lingual word embeddings by either joint training or post-training mappings of monolingual embeddings. Conneau et al. (2017) and Artetxe et al. (2018) first show two monolingual embeddings can be aligned by learning an orthogonal mapping with only identical strings as an initial heuristic bilingual dictionary.

Contextual Word Embeddings ELMo (Peters et al., 2018), a deep LSTM (Hochreiter and Schmidhuber, 1997) pretrained with a language modeling objective, learns contextual word embeddings. This contextualized representation outperforms stand-alone word embeddings, e.g. Word2Vec (Mikolov et al., 2013b) and Glove (Pennington et al., 2014), with the same task-specific architecture in various downstream tasks. Instead of taking the representation from a pretrained model, GPT (Radford et al., 2018) and Howard and Ruder (2018) also fine-tune all the parameters of the pretrained model for a specific task. Also, GPT uses a transformer encoder (Vaswani et al., 2017) instead of an LSTM and jointly fine-tunes with the language modeling objective. Howard and Ruder (2018) propose another fine-tuning strategy by using a different learning rate for each layer with learning rate warmup and gradual unfreezing.

Concurrent work by Lample and Conneau (2019) incorporates bitext into BERT by training on pairs of parallel sentences. Schuster et al. (2019) aligns pretrained ELMo of different languages by learning an orthogonal mapping and shows strong zero-shot and few-shot cross-lingual transfer performance on dependency parsing with 5 Indo-European languages. Similar to multilingual BERT, Mulcaire et al. (2019) trains a single ELMo on distantly related languages and shows mixed results as to the benefit of pretaining.

Parallel to our work, Pires et al. (2019) shows mBERT has good zero-shot cross-lingual transfer performance on NER and POS tagging. They show how subword overlap and word ordering effect mBERT transfer performance. Additionally, they show mBERT can find translation pairs and works on code-switched POS tagging. In comparison, our work looks at a larger set of NLP tasks including dependency parsing and ground the mBERT performance against previous state-of-the-art on zero-shot cross-lingual transfer. We also probe mBERT in different ways and show a more complete picture of the cross-lingual effectiveness of mBERT.

3 Multilingual BERT

BERT (Devlin et al., 2019) is a deep contextual representation based on a series of transformers trained by a self-supervised objective. One of the main differences between BERT and related work like ELMo and GPT is that BERT is trained by the Cloze task (Taylor, 1953), also referred to as masked language modeling, instead of right-to-left or left-to-right language modeling. This allows the model to freely encode information from both directions in each layer. Additionally, BERT also optimizes a next sentence classification objective. At training time, 50% of the paired sentences are consecutive sentences while the rest of the sentences are paired randomly. Instead of operating on words, BERT uses a subword vocabulary with WordPiece (Wu et al., 2016), a data-driven approach to break up a word into subwords.

Fine-tuning BERT BERT shows strong performance by fine-tuning the transformer encoder followed by a softmax classification layer on various sentence classification tasks. A sequence of shared softmax classifications produces sequence tagging models for tasks like NER. Fine-tuning usually takes 3 to 4 epochs with a relatively small learning rate, for example, 3e-5.

Multilingual BERT mBERT (Devlin, 2018) follows the same model architecture and training procedure as BERT, except with data from Wikipedia in 104 languages. Training makes no use of explicit cross-lingual signal, e.g. pairs of words, sentences or documents linked across languages. In mBERT, the WordPiece modeling strategy allows the model to share embeddings across languages. For example, "DNA" has a similar meaning even in distantly related languages like English and Chinese [1]. To account for varying sizes of Wikipedia training data in different languages, training uses a heuristic to subsample or oversample words when running WordPiece as well as sampling a training batch, random words for cloze and random sentences for next sentence classification.

Transformer For completeness, we describe the Transformer used by BERT. Let $\mathbf{x}$, $\mathbf{y}$ be a sequence of subwords from a sentence pair. A special token `[CLS]` is prepended to $\mathbf{x}$ and `[SEP]` is appended to both $\mathbf{x}$ and $\mathbf{y}$. The embedding is obtained by

$$\begin{aligned}\hat{h}_i^0 &= E(x_i) + E(i) + E(\mathbb{1}_{\mathbf{x}})\\ \hat{h}_{j+|\mathbf{x}|}^0 &= E(y_j) + E(j + |\mathbf{x}|) + E(\mathbb{1}_{\mathbf{y}})\\ h_\cdot^0 &= \text{Dropout}(\text{LN}(\hat{h}_\cdot^0))\end{aligned}$$

where E is the embedding function and LN is layer normalization (Ba et al., 2016). M transformer blocks are followed by the embeddings. In each transformer block,

$$\begin{aligned}h_\cdot^{i+1} &= \text{Skip}(\text{FF}, \text{Skip}(\text{MHSA}, h_\cdot^i))\\ \text{Skip}(f, h) &= \text{LN}(h + \text{Dropout}(f(h)))\\ \text{FF}(h) &= \text{GELU}(h\mathbf{W}_1^\top + \mathbf{b}_1)\mathbf{W}_2^\top + \mathbf{b}_2\end{aligned}$$

where GELU is an element-wise activation function (Hendrycks and Gimpel, 2016). In practice, $h^i \in \mathbb{R}^{(|\mathbf{x}|+|\mathbf{y}|)\times d_h}$, $\mathbf{W}_1 \in \mathbb{R}^{4d_h\times d_h}$, $\mathbf{b}_1 \in \mathbb{R}^{4d_h}$, $\mathbf{W}_2 \in \mathbb{R}^{d_h\times 4d_h}$, and $\mathbf{b}_2 \in \mathbb{R}^{d_h}$. MHSA is the multi-heads self-attention function. We show how one new position $\hat{h}_i$ is computed.

$$\begin{aligned}[\cdots, \hat{h}_i, \cdots] &= \text{MHSA}([h_1, \cdots, h_{|\mathbf{x}|+|\mathbf{y}|}])\\ &= \mathbf{W}_o\text{Concat}(h_i^1, \cdots, h_i^N) + \mathbf{b}_o\end{aligned}$$

In each attention, referred to as attention head,

$$\begin{aligned}h_i^j &= \sum_{k=1}^{|\mathbf{x}|+|\mathbf{y}|} \text{Dropout}(\alpha_k^{(i,j)})\mathbf{W}_V^j h_k\\ \alpha_k^{(i,j)} &= \frac{\exp \frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_k}{\sqrt{d_h/N}}}{\sum_{k'=1}^{|\mathbf{x}|+|\mathbf{y}|} \exp \frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_{k'}}{\sqrt{d_h/N}}}\end{aligned}$$

where N is the number of attention heads, $h_i^j \in \mathbb{R}^{d_h/N}$, $\mathbf{W}_o \in \mathbb{R}^{d_h\times d_h}$, $\mathbf{b}_o \in \mathbb{R}^{d_h}$, and $\mathbf{W}_Q^j, \mathbf{W}_K^j, \mathbf{W}_V^j \in \mathbb{R}^{d_h/N\times d_h}$.

4 Tasks

Does mBERT learn a cross-lingual representation, or does it produce a representation for each language in its own embedding space? We consider five tasks in the zero-shot transfer setting. We assume labeled training data for each task in English, and transfer the trained model to a target language. We select a range of different tasks: document classification, natural language inference, named entity recognition, part-of-speech tagging, and dependency parsing. We cover zero-shot transfer from English to 38 languages in the 5 different tasks as shown in Tab. 1. In this section, we describe the tasks as well as task-specific layers.

4.1 Document Classification

We use MLDoc (Schwenk and Li, 2018), a balanced subset of the Reuters corpus covering 8 languages for document classification. The 4-way topic classification task decides between CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). We only use the first two sentences[2] of a document for classification due to memory constraint. The sentence pairs are provided to the mBERT encoder. The task-specific classification layer is a linear function mapping $h_0^{12} \in \mathbb{R}_h^d$ into $\mathbb{R}^4$, and a softmax is used to get class distribution. We evaluate by classification accuracy.

4.2 Natural Language Inference

We use XNLI (Conneau et al., 2018) which cover 15 languages for natural language inference. The 3-way classification includes entailment, neutral, and contradiction given a pair of sentences. We

[1] "DNA" indeed appears in the vocabulary of mBERT as a stand-alone lexicon.

[2] We only use the first sentence if the document only contains one sentence. Documents are segmented into sentences with NLTK (Perkins, 2014).

	ar	bg	ca	cs	da	de	el	en	es	et	fa	fi	fr	he	hi	hr	hu	id	it	ja	ko	la	lv	nl	no	pl	pt	ro	ru	sk	sl	sv	sw	th	tr	uk	ur	vi	zh
MLDoc						✓		✓	✓				✓						✓	✓									✓										✓
NLI	✓	✓				✓	✓	✓	✓				✓		✓														✓				✓	✓	✓		✓	✓	✓
NER						✓		✓	✓															✓															✓
POS		✓			✓	✓		✓	✓		✓						✓		✓					✓		✓	✓	✓		✓	✓	✓							
Parsing	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓			✓

Table 1: The 39 languages used in the 5 tasks.

feed a pair of sentences directly into mBERT and the task-specific classification layer is the same as §4.1. We evaluate by classification accuracy.

4.3 Named Entity Recognition

We use the CoNLL 2002 and 2003 NER shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) (4 languages) and a Chinese NER dataset (Levow, 2006). The labeling scheme is BIO with 4 types of named entities. We add a linear classification layer with softmax to obtain word-level predictions. Since mBERT operates at the subword-level while the labeling is word-level, if a word is broken into multiple subwords, we mask the prediction of non-first subwords. NER is evaluated by F1 of predicted entity (F1). Note we use a simple post-processing heuristic to obtain a valid span.

4.4 Part-of-Speech Tagging

We use a subset of Universal Dependencies (UD) Treebanks (v1.4) (Nivre et al., 2016), which cover 15 languages, following the setup of Kim et al. (2017). The task-specific labeling layer is the same as §4.3. POS tagging is evaluated by the accuracy of predicted POS tags (ACC).

4.5 Dependency parsing

Following the setup of Ahmad et al. (2019), we use a subset of Universal Dependencies (UD) Treebanks (v2.2) (Nivre et al., 2018), which includes 31 languages. Dependency parsing is evaluated by unlabelled attachment score (UAS) and labeled attachment score (LAS) [3]. We only predict the coarse-grain dependency label following Ahmad et al. We use the model of Dozat and Manning (2016), a graph-based parser as a task-specific layer. Their LSTM encoder is replaced by mBERT. Similar to §4.3, we only take the representation of the first subword of each word. We use masking to prevent the parser from operating on non-first subwords.

[3] Punctuations (PUNCT) and symbols (SYM) are excluded.

5 Experiments

We use the base cased multilingual BERT, which has $N = 12$ attention heads and $M = 12$ transformer blocks. The dropout probability is 0.1 and d_h is 768. The model has 179M parameters with about 120k vocabulary.

Training For each task, no preprocessing is performed except tokenization of words into subwords with WordPiece. We use Adam (Kingma and Ba, 2014) for fine-tuning with β_1 of 0.9, β_2 of 0.999 and L2 weight decay of 0.01. We warm up the learning rate over the first 10% of batches and linearly decay the learning rate.

Maximum Subwords Sequence Length At training time, we limit the length of subwords sequence to 128 to fit in a single GPU for all tasks. For NER and POS tagging, we additionally use the sliding window approach. After the first window, we keep the last 64 subwords from the previous window as context. In other words, for a non-first window, only (up to) 64 new subwords are added for prediction. At evaluation time, we follow the same approach as training time except for parsing. We threshold the sentence length to 140 words, including words and punctuation, following Ahmad et al. (2019). In practice, the maximum subwords sequence length is the number of subwords of the first 140 words or 512, whichever is smaller.

Hyperparameter Search and Model Selection We select the best hyperparameters by searching a combination of batch size, learning rate and the number of fine-tuning epochs with the following range: learning rate $\{2\times10^{-5}, 3\times10^{-5}, 5\times10^{-5}\}$; batch size $\{16, 32\}$; number of epochs: $\{3, 4\}$. Note the best hyperparameters and model are selected by development performance in *English*.

5.1 Question #1: Is mBERT Multilingual?

MLDoc We include two strong baselines. Schwenk and Li (2018) use MultiCCA, multilingual word embeddings trained with a bilingual dictionary (Ammar et al., 2016), and convolution neural networks. Concurrent to our work,

	en	de	zh	es	fr	it	ja	ru	Average
In language supervised learning									
Schwenk and Li (2018)	92.2	93.7	87.3	94.5	92.1	85.6	85.4	85.7	89.5
mBERT	94.2	93.3	89.3	95.7	93.4	88.0	88.4	87.5	91.2
Zero-shot cross-lingual transfer									
Schwenk and Li (2018)	<u>92.2</u>	<u>81.2</u>	<u>74.7</u>	72.5	72.4	**69.4**	**67.6**	60.8	73.9
Artetxe and Schwenk (2018) ♠ †	89.9	**84.8**	71.9	**77.3**	**78.0**	**69.4**	<u>60.3</u>	<u>67.8</u>	**74.9**
mBERT	**94.2**	80.2	**76.9**	<u>72.6</u>	<u>72.6</u>	<u>68.9</u>	56.5	**73.7**	<u>74.5</u>

Table 2: MLDoc experiments. ♠ denotes the model is pretrained with bitext, and † denotes concurrent work. Bold and underline denote best and second best.

Artetxe and Schwenk (2018) use bitext between English/Spanish and the rest of languages to pretrain a multilingual sentence representation with a sequence-to-sequence model where the decoder only has access to a max-pooling of the encoder hidden states.

mBERT outperforms (Tab. 2) multilingual word embeddings and performs comparably with a multilingual sentence representation, even though mBERT does not have access to bitext. Interestingly, mBERT outperforms Artetxe and Schwenk (2018) in distantly related languages like Chinese and Russian and under-performs in closely related Indo-European languages.

XNLI We include three strong baselines, Artetxe and Schwenk (2018) and Lample and Conneau (2019) are concurrent to our work. Lample and Conneau (2019) with MLM is similar to mBERT; the main difference is that it only trains with the 15 languages of XNLI, has 249M parameters (around 40% more than mBERT), and MLM+TLM also uses bitext as training data [4]. Conneau et al. (2018) use supervised multilingual word embeddings with an LSTM encoder and max-pooling. After an English encoder and classifier are trained, the target encoder is trained to mimic the English encoder with ranking loss and bitext.

In Tab. 3, mBERT outperforms one model with bitext training but (as expected) falls short of models with more cross-lingual training information. Interestingly, mBERT and MLM are mostly the same except for the training languages, yet we observe that mBERT under-performs MLM by a large margin. We hypothesize that limiting pretraining to only those languages needed for the downstream task is beneficial. The gap between Artetxe and Schwenk (2018) and mBERT in XNLI is larger than MLDoc, likely because XNLI is harder.

[4]They also use language embeddings as input and exclude the next sentence classification objective

NER We use Xie et al. (2018) as a zero-shot cross-lingual transfer baseline, which is state-of-the-art on CoNLL 2002 and 2003. It uses unsupervised bilingual word embeddings (Conneau et al., 2017) with a hybrid of a character-level/word-level LSTM, self-attention, and a CRF. Pseudo training data is built by word-to-word translation with an induced dictionary from bilingual word embeddings.

mBERT outperforms a strong baseline by an average of 6.9 points absolute F1 and an 11.8 point absolute improvement in German with a simple one layer 0^{th}-order CRF as a prediction function (Tab. 4). A large gap remains when transferring to distantly related languages (e.g. Chinese) compared to a supervised baseline. Further effort should focus on transferring between distantly related languages. In §5.4 we show that sharing subwords across languages helps transfer.

POS We use Kim et al. (2017) as a reference. They utilized a small amount of supervision in the target language as well as English supervision so the results are not directly comparable. Tab. 5 shows a large (average) gap between mBERT and Kim et al. Interestingly, mBERT still outperforms Kim et al. (2017) with 320 sentences in German (de), Polish (pl), Slovak (sk) and Swedish (sv).

Dependency Parsing We use the best performing model on average in Ahmad et al. (2019) as a zero-shot transfer baseline, i.e. transformer encoder with graph-based parser (Dozat and Manning, 2016), and dictionary supervised cross-lingual embeddings (Smith et al., 2017). Dependency parsers, including Ahmad et al., assume access to gold POS tags: a cross-lingual representation. We consider two versions of mBERT: with and without gold POS tags. When tags are available, a tag embedding is concatenated with the final output of mBERT.

Tab. 6 shows that mBERT outperforms the base-

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Average
Pseudo supervision with machine translated training data from English to target language																
Lample and Conneau (2019) (MLM+TLM) ♠ †	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
mBERT	82.1	76.9	78.5	74.8	72.1	75.4	74.3	70.6	70.8	67.8	63.2	76.2	65.3	65.3	60.6	71.6
Zero-shot cross-lingual transfer																
Conneau et al. (2018) (X-LSTM) ♠ ♢	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Artetxe and Schwenk (2018) ♠ †	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
Lample and Conneau (2019) (MLM+TLM) ♠ ♢ †	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Lample and Conneau (2019) (MLM) ♢ †	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
mBERT	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3

Table 3: XNLI experiments. ♠ denotes the model is pretrained with cross-lingual signal including bitext or bilingual dictionary, † denotes concurrent work, and ♢ denotes model selection with target language dev set.

	en	nl	es	de	zh	Average (-en,-zh)
In language supervised learning						
Xie et al. (2018)	-	86.40	86.26	78.16	-	83.61
mBERT	91.97	90.94	87.38	82.82	93.17	87.05
Zero-shot cross-lingual transfer						
Xie et al. (2018)	-	71.25	72.37	57.76	-	67.13
mBERT	91.97	**77.57**	**74.96**	**69.56**	51.90	**74.03**

Table 4: NER tagging experiments.

line on average by 7.3 point UAS and 0.4 point LAS absolute improvement even without gold POS tags. Note in practice, gold POS tags are not always available, especially for low resource languages. Interestingly, the LAS of mBERT tends to weaker than the baseline in languages with less word order distance, in other words, more closely related to English. With the help of gold POS tags, we further observe 1.6 points UAS and 4.7 point LAS absolute improvement on average. It appears that adding gold POS tags, which provide clearer cross-lingual representations, benefit mBERT.

Summary Across all five tasks, mBERT demonstrate strong (sometimes state-of-the-art) zero-shot cross-lingual performance without any cross-lingual signal. It outperforms cross-lingual embeddings in four tasks. With a small amount of target language supervision and cross-lingual signal, mBERT may improve further; we leave this as future work. In short, mBERT is a surprisingly effective cross-lingual model for many NLP tasks.

5.2 Question #2: Does mBERT vary layer-wise?

The goal of a deep neural network is to abstract to higher-order representations as you progress up the hierarchy (Yosinski et al., 2014). Peters et al. (2018) empirically show that for ELMo in English the lower layer is better at syntax while the upper layer is better at semantics. However, it is unclear how different layers affect the quality of cross-lingual representation. For mBERT, we hypothesize a similar generalization across the 13 layers, as well as an abstraction away from a specific language with higher layers. Does the zero-shot transfer performance vary with different layers?

We consider two schemes. First, we follow the feature-based approach of ELMo by taking a learned weighted combination of all 13 layers of mBERT with a two-layer bidirectional LSTM with d_h hidden size (Feat). Note the LSTM is trained from scratch and mBERT is fixed. For sentence and document classification, an additional max-pooling is used to extract a fixed-dimension vector. We train the feature-based approach with Adam and learning rate 1e-3. The batch size is 32. The learning rate is halved whenever the development evaluation does not improve. The training is stopped early when learning rate drop below 1e-5. Second, when fine-tuning mBERT, we fix the bottom n layers (n included) of mBERT, where layer 0 is the input embedding. We consider $n \in \{0, 3, 6, 9\}$.

Freezing the bottom layers of mBERT, in general, improves the performance of mBERT in all five tasks (Fig. 1). For sentence-level tasks like document classification and natural language inference, we observe the largest improvement with $n = 6$. For word-level tasks like NER, POS tagging, and parsing, we observe the largest improvement with $n = 3$. More improvement in under-performing languages is observed.

In each task, the feature-based approach with LSTM under-performs fine-tuning approach. We hypothesize that initialization from pretraining with lots of languages provides a very good starting point that is hard to beat. Additionally, the LSTM could also be part of the problem. In Ahmad et al. (2019) for dependency parsing, an LSTM encoder was worse than a transformer when transferring

lang	bg	da	de	en	es	fa	hu	it	nl	pl	pt	ro	sk	sl	sv	Average (-en)
In language supervised learning																
mBERT	99.0	97.9	95.2	97.1	97.1	97.8	96.9	98.7	92.1	98.5	98.3	97.8	97.0	98.9	98.4	97.4
Low resource cross-lingual transfer																
Kim et al. (2017) (1280)	95.7	94.3	90.7	-	93.4	94.8	94.5	95.9	85.8	92.1	95.5	94.2	90.0	94.1	94.6	93.3
Kim et al. (2017) (320)	92.4	90.8	89.7	-	90.9	91.8	90.7	94.0	82.2	85.5	94.2	91.4	83.2	90.6	90.7	89.9
Zero-shot cross-lingual transfer																
mBERT	87.4	88.3	89.8	97.1	85.2	72.8	83.2	84.7	75.9	86.9	82.1	84.7	83.6	84.2	91.3	84.3

Table 5: POS tagging. Kim et al. (2017) use small amounts of training data in the target language.

	Dist	mBERT(S)	Baseline(Z)	mBERT(Z)	mBERT(Z+POS)
en	0.00	91.5/81.3	90.4/**88.4**	<u>91.5</u>/81.3	**91.8**/<u>82.2</u>
no	0.06	93.6/85.9	<u>80.8</u>/**72.8**	80.6/68.9	**82.7**/<u>72.1</u>
sv	0.07	91.2/83.1	81.0/<u>73.2</u>	<u>82.5</u>/71.2	**84.3/73.7**
fr	0.09	91.7/85.4	77.9/<u>72.8</u>	<u>82.7</u>/72.7	**83.8/76.2**
pt	0.09	93.2/87.2	76.6/**67.8**	<u>77.1</u>/64.0	**78.3**/<u>66.9</u>
da	0.10	89.5/81.9	76.6/<u>67.9</u>	<u>77.4</u>/64.7	**79.3/68.1**
es	0.12	92.3/86.5	74.5/<u>66.4</u>	<u>78.1</u>/64.9	**79.0/68.9**
it	0.12	94.8/88.7	80.8/<u>75.8</u>	<u>84.6</u>/74.4	**86.0/77.8**
ca	0.13	94.3/89.5	73.8/<u>65.1</u>	<u>78.1</u>/64.6	**79.0/67.9**
hr	0.13	92.4/83.8	61.9/52.9	**80.7**/<u>65.8</u>	<u>80.4</u>/**68.2**
pl	0.13	94.7/79.9	74.6/<u>62.2</u>	<u>82.8</u>/59.4	**85.7/65.4**
sl	0.13	88.0/77.8	68.2/<u>56.5</u>	<u>72.6</u>/51.4	**75.9/59.2**
uk	0.13	90.6/83.4	60.1/52.3	**76.7**/<u>60.0</u>	<u>76.5</u>/**65.5**
bg	0.14	95.2/85.5	79.4/**68.2**	<u>83.3</u>/62.3	**84.4**/<u>68.1</u>
cs	0.14	94.2/86.6	63.1/53.8	<u>76.6</u>/<u>58.7</u>	**77.4/63.6**
de	0.14	86.1/76.5	71.3/61.6	<u>80.4</u>/<u>66.3</u>	**83.5/71.2**
he	0.14	91.9/83.6	55.3/48.0	**67.5**/<u>48.4</u>	<u>67.0</u>/**54.3**
nl	0.14	94.0/85.0	68.6/60.3	<u>78.0</u>/<u>64.8</u>	**79.9/67.1**
ru	0.14	94.7/88.0	60.6/51.6	**73.6**/<u>58.5</u>	<u>73.2</u>/**61.5**
ro	0.15	92.2/83.2	65.1/54.1	**77.0**/<u>58.5</u>	<u>76.9</u>/**62.6**
id	0.17	86.3/75.4	49.2/43.5	**62.6**/<u>45.6</u>	<u>59.8</u>/**48.6**
sk	0.17	93.8/83.3	66.7/58.2	<u>82.7</u>/<u>63.9</u>	**82.9/67.8**
lv	0.18	87.3/75.3	**70.8/49.3**	66.0/41.4	<u>70.4</u>/<u>48.5</u>
et	0.20	88.8/79.7	65.7/<u>44.9</u>	<u>66.9</u>/44.3	**70.8/50.7**
fi	0.20	91.3/81.8	66.3/<u>48.7</u>	<u>68.4</u>/47.5	**71.4/52.5**
zh*	0.23	88.3/81.2	42.5/25.1	**53.8**/<u>26.8</u>	<u>53.4</u>/**29.0**
ar	0.26	87.6/80.6	38.1/28.0	<u>43.9</u>/<u>28.3</u>	**44.7/32.9**
la	0.28	85.2/73.1	<u>48.0</u>/**35.2**	47.9/26.1	**50.9**/<u>32.2</u>
ko	0.33	86.0/74.8	34.5/16.4	**52.7**/<u>27.5</u>	<u>52.3</u>/**29.4**
hi	0.40	94.8/86.7	35.5/26.5	<u>49.8</u>/<u>33.2</u>	**58.9/44.0**
ja*	0.49	94.2/87.4	28.2/<u>20.9</u>	<u>36.6</u>/15.7	**41.3/30.9**
AVER	0.17	91.3/82.6	64.1/53.8	<u>71.4</u>/<u>54.2</u>	**73.0/58.9**

Table 6: Dependency parsing results by language (UAS/LAS). * denotes delexicalized parsing in the baseline. S and Z denotes supervised learning and zero-shot transfer. Bold and underline denotes best and second best. We order the languages by word order distance to English.

to languages with high word ordering distance to English.

5.3 Question #3: Does mBERT retain language specific information?

mBERT may learn a cross-lingual representation by abstracting away from language-specific information, thus losing the ability to distinguish between languages. We test this by considering language identification: does mBERT retain language-specific information? We use WiLI-2018 (Thoma, 2018), which includes over 200 languages from Wikipedia. We keep only those languages included in mBERT, leaving 99 languages [5]. We take various layers of bag-of-words mBERT representation of the first two sentences of the test paragraph and add a linear classifier with softmax. We fix mBERT and train *only* the classifier the same as the feature-based approach in §5.2.

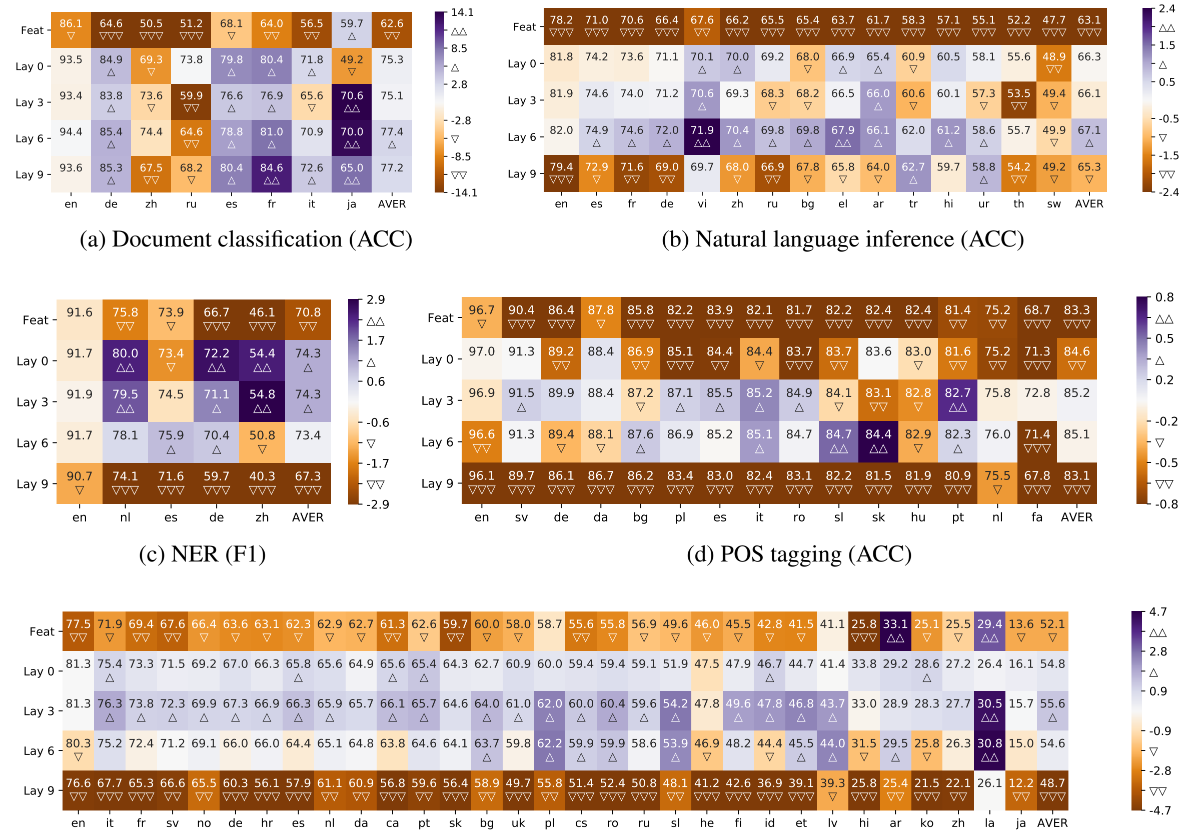
All tested layers achieved around 96% accuracy (Fig. 2), with no clear difference between layers. This suggests each layer contains language-specific information; surprising given the zero-shot cross-lingual abilities. As mBERT generalizes its representations and creates cross-lingual representations, it maintains language-specific details. This may be encouraged during pretraining since mBERT needs to retain enough language-specific information to perform the cloze task.

5.4 Question #4: Does mBERT benefit by sharing subwords across languages?

As discussed in §3, mBERT shares subwords in closely related languages or perhaps in distantly related languages. At training time, the representation of a shared subword is explicitly trained to contain enough information for the cloze task in all languages in which it appears. During fine-tuning for zero-shot cross-lingual transfer, if a subword in the target language test set also appears in the source language training data, the supervision could be leaked to the target language explicitly. However, all subwords interact in a non-interpretable way inside a deep network, and subword representations could overfit to the source language and potentially hurt transfer performance. In these experiments, we investigate how sharing subwords across languages effects cross-lingual transfer.

To quantify how many subwords are shared

[5] Hungarian, Western-Punjabi, Norwegian-Bokmal, and Piedmontese are not covered by WiLI.



(a) Document classification (ACC)

(b) Natural language inference (ACC)

(c) NER (F1)

(d) POS tagging (ACC)

(e) Dependency parsing (LAS)

Figure 1: Performance of different fine-tuning approaches compared with fine-tuning all mBERT parameters. Color denotes absolute difference and number in each entry is the evaluation in the corresponding setting. Languages are sorted by mBERT zero-shot transfer performance. Three downward triangles indicate performance drop more than the legends lower limit.

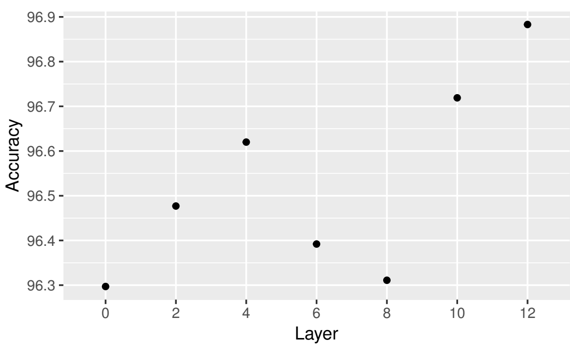


Figure 2: Language identification accuracy for different layer of mBERT. layer 0 is the embedding layer and the layer $i > 0$ is output of the i[th] transformer block.

across languages in any task, we assume $V^{\text{en}}_{\text{train}}$ is the set of all subwords in the English training set, V^{ℓ}_{test} is the set of all subwords in language ℓ test set, and c^{ℓ}_{w} is the count of subword w in test set of language ℓ. We then calculate the percentage of observed subwords at type-level p^{ℓ}_{type} and token-level p^{ℓ}_{token} for each target language ℓ.

$$p^{\ell}_{\text{type}} = \frac{|V^{\ell}_{\text{obs}}|}{|V^{\ell}_{\text{test}}|} \times 100$$

$$p^{\ell}_{\text{token}} = \frac{\sum_{w \in V^{\ell}_{\text{obs}}} c^{\ell}_{w}}{\sum_{w \in V^{\ell}_{\text{test}}} c^{\ell}_{w}} \times 100$$

where $V^{\ell}_{\text{obs}} = V^{\text{en}}_{\text{train}} \cap V^{\ell}_{\text{test}}$.

In Fig. 3, we show the relation between cross-lingual zero-shot transfer performance of mBERT and p^{ℓ}_{type} or p^{ℓ}_{token} for all five tasks with Pearson correlation. In four out of five tasks (not XNLI) we observed a strong positive correlation ($p < 0.05$) with a correlation coefficient larger than 0.5. In Indo-European languages, we observed p^{ℓ}_{token} is usually around 50% to 75% while p^{ℓ}_{type} is usually less than 50%. This indicates that subwords shared across languages are usually high frequency[6]. We

[6]With the data-dependent WordPiece algorithm, subwords that appear in multiple languages with high frequency are more likely to be selected.

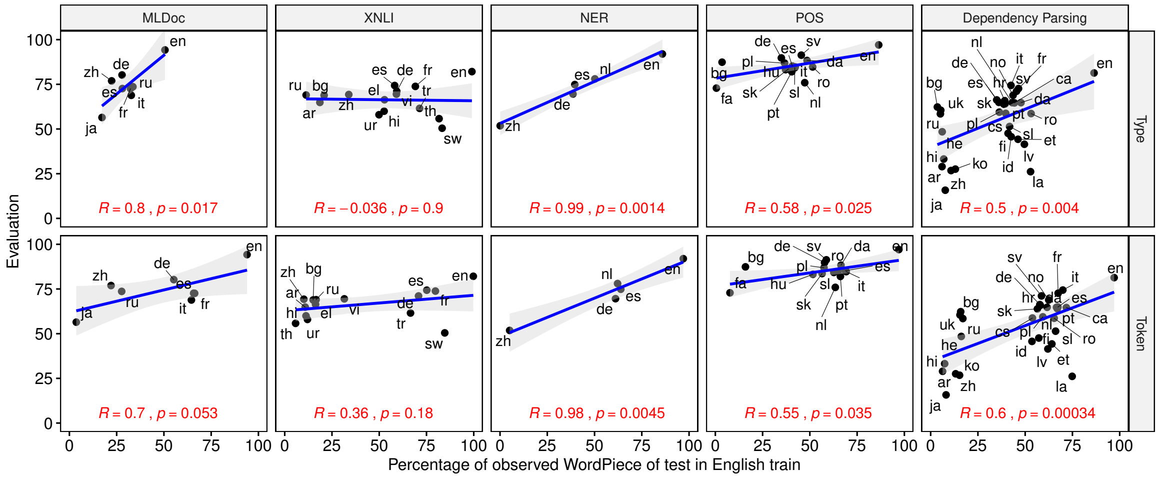


Figure 3: Relation between cross-lingual zero-shot transfer performance with mBERT and percentage of observed subwords at both type-level and token-level. Pearson correlation coefficient and p-value are shown in red.

hypothesize that this could be used as a simple indicator for selecting source language in cross-lingual transfer with mBERT. We leave this for future work.

6 Discussion

We show mBERT does well in a cross-lingual zero-shot transfer setting on five different tasks covering a large number of languages. It outperforms cross-lingual embeddings, which typically have more cross-lingual supervision. By fixing the bottom layers of mBERT during fine-tuning, we observe further performance gains. Language-specific information is preserved in all layers. Sharing subwords helps cross-lingual transfer; a strong correlation is observed between the percentage of overlapping subwords and transfer performance.

mBERT effectively learns a good multilingual representation with strong cross-lingual zero-shot transfer performance in various tasks. We recommend building future multi-lingual NLP models on top of mBERT or other models pretrained similarly. Even without explicit cross-lingual supervision, these models do very well. As we show with XNLI in §5.1, while bitext is hard to obtain in low resource settings, a variant of mBERT pretrained with bitext (Lample and Conneau, 2019) shows even stronger performance. Future work could investigate how to use weak supervision to produce a better cross-lingual mBERT, or adapt an already trained model for cross-lingual use. With POS tagging in §5.1, we show mBERT, in general, under-performs models with a small amount of supervision while Devlin et al. (2019) show that in English NLP tasks, fine-tuning BERT only needs a small amount of data. Future work could investigate when cross-lingual transfer is helpful in NLP tasks of low resource languages. With such strong cross-lingual NLP performance, it would be interesting to prob mBERT from a linguistic perspective in the future.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin. 2018. Multilingual bert readme document.

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.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. 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 3912–3918, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Phng Lê H`ông, 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, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ 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 Roca, 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.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phng Lê H`ông, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2016. Universal dependencies 1.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jacob Perkins. 2014. *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. 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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

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

Martin Thoma. 2018. The wili benchmark dataset for written language identification. *arXiv preprint arXiv:1801.07779*.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

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*.