Does Multilingual BERT Have an English Accent?

Stanford CS224N Custom Project

Kezia Lopez
Department of Computer Science
Stanford University
keziakl@stanford.edu

Abstract

Multilingual BERT is a powerful tool to perform language learning transfer tasks, especially for low-resource languages. However, the training data M-BERT uses varies greatly in quantity and quality depending on the language. We explore whether this bias in training data makes M-BERT produce "english-esque" results on the task of sentence classification and rating for pronoun-dropped sentences in Spanish, after being fine-tuned to the task in the target language. We also make a deeper comparison by having the models "rate" pro-drop sentences to see which of the models consistently "likes" the pro-drop sentences more.

1 Key Information to include

- Mentor: Vincent Li, Isabel Papadimitriou
- External Collaborators (if you have any): Isabel Papadimitriou, Prof. Dan Jurafsky
- Sharing project: No

2 Introduction

Native or Heritage speakers of non-English languages in the United States experience language influence, whether on their grammar or lexicon. Heritage Spanish speakers, particularly exhibited by Puerto Rican speakers of "Spanglish" and Chicanx (Americans of Mexican origin or descent) Spanish speakers, exhibit many influences of English on their word choice and phrasal structures, such as English-like tense use and word order. When a speaker's predominant language they interact in their daily life is not their native or heritage tongue, the native or heritage language may experience psycholinguistical shifts in the representation of grammar or semantics. We plan to replicate this real-world linguistic phenomena using a computer to answer the following: How well does Multilingual BERT, a BERT model trained on over a hundred languages, represent languages other than English, and will Multilingual BERT, whose largest training set is English Wikipedia, be able to successfully perform tasks in Spanish without suffering influences of English and producing English-esque results?

Multilingual BERT (mBERT) is a powerful language model trained on a concatenated monolingual Wikipedia corpora in 104 languages [1]. It does not use any markers to denote the input language when training and does not explicitly encourage translation pairs to have similar representations, so it is fascinating that mBERT performs state-of-the-art in many language tasks.

Researchers have created many non-English monolingual models, such as BETO for Spanish and FinBERT for Finnish, that copy the traditional BERT masked-language-modelling structure. These large models are costly to train and only perform well for relatively high-resource languages (languages with lots of written data able to be parsed efficiently by a computer). The researchers that introduce mBERT promise that their model can help solve low-resource language data scarcity by serving as a base model that people can fine-tune to their target language task.

We put mBERT to the test in Spanish, by comparing it to BETO, a monolingual Spanish model that performs state-of-the-art in many tasks. mBERT is particularly well-suited for a probing study because it allows for straightforward zero-shot cross-lingual model transfer. We first use mBERT and BETO to classify a sequence (sentence) as grammatically correct or incorrect after finetuning on a small corpus of labelled Spanish text. Then, to delve deeper in a second experiment, we use mBERT and BETO to assign each sentence with a "score" by randomly sampling the logits (the log probability of a token being itself given its context) of the k tokens, adding all k to get a pseudo-chain-rule probability, normalizing with a softmax to get a readable result. We find that BETO "likes" pro-drop sentences more than BERT does, even though there is no question about their grammaticality.

3 Related Work

We use what Google researchers' Pires et al. show in their publication of "How Multilingual is Multilingual BERT?" to gain inspiration into language model probing tasks. According to their paper, Multilingual BERT is able to perform cross-lingual generalization surprisingly well. Although high lexical overlap between languages improves transfer, mBERT is even able to transfer between languages that are written in completely different scripts– this mean ZERO lexical overlap– which shows that mBERT is able to capture multilingual representations! However, although mBERT is able to transfer learning for typologically similar languages (languages that share many features, such as adjective-noun order or grammatical cases) and map learned structures on these new language vocabularies, it does not seem to learn the underlying transformations of syntactic structures (the formation of language syntax trees) to accommodate target languages with different word order.

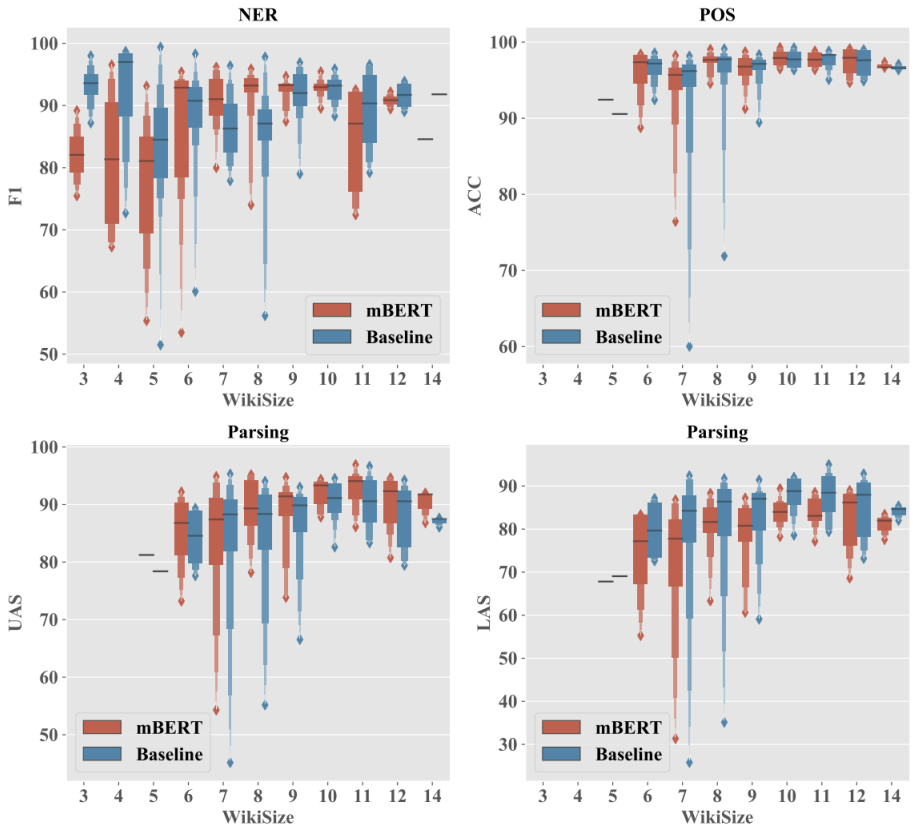

Figure 1: mBERT vs baseline grouped by WikiSize. mBERT performance drops much more than baseline models on languages lower than WikiSize 6 – the bottom 30% languages supported by mBERT – especially in NER, which covers nearly all mBERT supported languages. Breakdown can be found in App. A.

Many papers explore mBERT's ability to generalize, its language neutrality, and its language-

knowledge transfer capabilities. Of course, Pires et al. show that mBERT is "surprisingly good" at zero-shot cross-lingual transfer in several European languages and high-resource Asian languages with their public release of the model in 2019. A few months later, Libovicky et al. explore mBERT's "language neutrality" in their paper "How Language Neutral is Multilingual BERT?" [2]. Because Pires emphasized morphological and syntactical tasks, tasks like lexical and grammatical transfer, Libovicky and their team focus on the semantic properties of mBERT. They find that in the task of language identification (having mBERT identify what language a sentence is in), centering the sentence representations in each language so that their average lies at the centroid of the mBERT representation vector space considerably decreases the accuracy. This supports their claim that mBERT has a language-specific and language-neutral component to sentence representation. In their publication "Are All Languages Created Equal in Multilingual BERT", Wu et al. show that for languages with a WikiSize less than 6 (the bottom 30% of languages in terms of quality and quantity of data), mBERT goes from being competitive with state-of-the-art to being over 10 points behind for NER tasks [3]. This is shown in Figure 1 above.

In "Multilingual is not Enough: BERT for Finnish" and "Spanish Pre-Trained BERT Model and Evaluation Data", Virtanen et al. and Canete et al. show that even though their monolingual models were not trained on as big of a dataset as mBERT, they perform notably better for most tasks like POS-tagging, NER, and dependency parsing [4] [5]. These papers compare the monolingual models to mBERT in reference tasks; however, in this paper, we delve deeper into specific areas of typology that mBERT may incorrectly generalize.

4 Approach

Pronoun dropping is a common occurrence in a lot of languages, including Spanish. Pronoun dropping occurs when a speaker or writer drops the optional pronoun, usually the subject, and still creates a grammatically acceptable sentence. In English, this is not possible. For example, a sentence like "I ate yesterday." is grammatically acceptable in English, while a sentence with the pronoun dropped: "Ate yesterday" is not. In Spanish, however, the equivalent sentences "Yo comí ayer" and "Comí ayer" are both grammatically acceptable, and in most cases, the form with the pronoun dropped is actually preferred.

Our first is to test Multilingual BERT on the task of rating Spanish sentences with pronoun dropping (0 or 1), after being finetuned to the task of rating the "grammatical acceptability" of a sentence (with a 0 or 1, 0 being unacceptable, 1 being acceptable). We use BertForSequenceClassification, which is the BERT-base model trained using Masked Language Modelling, with an additional linear layer on top that condenses the tensor output to a single value per sentence. We not only measure the validation accuracy on the task, but also the validation loss, to see if, even when achieving correct results, which of the models has more trouble with the task.

As a second experiment, we have both of the models (without fine-tuning) assign each of those pronoun-dropped sentences with a "score" by taking k=10 samples of the log probability that mBERT and BETO assign to a token after softmax regularization and add all k to get a pseudo-chain-rule probability. We average this over all of the sentences to get a score between 0 and 1, 1 being the optimal probability, 0 being worst.

5 Experiments

A more detailed explanation of our experiments is below:

5.1 Data

Using the UD Treebanks dataset of Spanish, I concatenated all of the training, dev, and test sets into one document. Then, I ran a parser I created to find sentences that exhibited pronoun dropping. This created a dataset of well over 2000 sentences. I used the following to identify pronoun-dropped sentences.

- Make sure that the ROOT of the sentence is a verb as some of the "sentences" in the dataset are actually noun phrases without a verb.
- Get rid of sentences that use the verb "haber". In Spanish, the verb "haber" can be used as a copula to mean "it exists"; for example, in the English sentence "There are worms", the word "there" is not a "subject" that performs an action. I removed these sentences to reduce bias in the predictions as it is impossible for there to be a subject in these sentences.
- Remove sentences that have a subject (noun or pronoun). This includes passive sentences.
- Spanish has a special case of the passive that does not allow for a subject in certain verbs. To make sure that the models were not using confounding information to make decisions, I removed these sentences by removing those that contained the 'iobj' tag.

5.2 Evaluation method

For the first experiment, we use Chris McCormick's evaluation metrics (link here), which is a simple validation accuracy (correct prediction is plus one point, incorrect is plus zero points, then divide by the total number of predictions). I also record the training and validation losses after each epoch for both models.

For the second experiment, evaluating is trickier. Because it is non-intuitive to assign a probability for an entire sequence of tokens at once with BERT as it is a masked-language model that assigns probabilities with a token-by-token basis, we decided to use a chain-rule probability metric. We add the logits (log probabilities, the equivalent of multiplying probabilities) of a random sample. We use k=5 and k=10, and because the probabilities per token are relatively small (mBERT and BETO have vocabularies of 105,879 and 31,002 respectively), k=10, a larger sample, provides a better and more accurate score for a sentence. As a note, there were a minimal amount of sentences that contained less than 10 tokens. We did not count those in our average score calculation as because of the nature of probability multiplication, did not want to bias our results with significantly larger results just because of smaller token samples. We take the average of all of the scores added together for both mBERT and BETO and compare them.

5.3 Experimental details

For the first experiment, we used the learning rate and epsilon values for the Adam Optimizer used in the Hugging Face run_glue.py file (lr = 2e-5, eps=1e-8). We the bert-base-uncased version of the BertForSequenceClassification implementations for mBERT and BERT. We ran the models for 4 epochs each and recorded the validation accuracy and loss at each stage.

For the second experiment, we used the bert-base-uncased version of the BertForMaskedLM implementations for mBERT and BERT. We run the experiment once over the sentence dataset and record the average scores for both models. We do not finetune the models beforehand.

5.4 Results

We find that BETO systematically "likes" pronoun-dropped sentences better than mBERT does.

- For the first experiment, we receive the following results:
  Results for BETO Spanish model:

```
Training complete!
Total training took 0:03:07 (h:mm:ss)
       Training Loss  Valid. Loss  Valid. Accur. Training Time Validation Time
epoch
1          3.720e-02    2.192e-04            1.0       0:00:45         0:00:02
2          2.208e-04    1.307e-04            1.0       0:00:45         0:00:02
3          1.528e-04    1.022e-04            1.0       0:00:45         0:00:02
4          1.318e-04    9.445e-05            1.0       0:00:45         0:00:02
```

Results for MBERT model:

```
Training complete!
Total training took 0:03:07 (h:mm:ss)
       Training Loss  Valid. Loss  Valid. Accur. Training Time Validation Time
epoch
1          5.547e-02    6.148e-04            1.0       0:00:45         0:00:02
2          6.063e-04    3.449e-04            1.0       0:00:45         0:00:02
3          3.991e-04    2.590e-04            1.0       0:00:45         0:00:02
4          3.309e-04    2.359e-04            1.0       0:00:45         0:00:02
```

- As you can see, the vaildation accuracy is perfect for both models. This makes sense as we only trained on correct pro-drop sentences, and these transformer models perform very well on binary classification tasks. We generated incorrect training data using the strategies outlined in "Corpora Generation for Grammatical Error Correction" (link here) published in the ACL proceedings in 2019 [6]; however, ran into some issues and did not have enough time to implement the experiment using this dataset. Given more time, we can perform more experiments using not only this incorrect dataset, but also non-pro drop sentences as well. Regardless of the results of the validation accuracy, however, we see a very interesting trend in the loss of both models. Although both models reduce the training and validation loss with every epoch, BETO's training and validation loss is significantly lower and continues to decrease by several orders of magnitude in every single epoch, showing that it has an easier time rating these sentences as grammatically correct.
- For the second experiment, we record the final average cumulative probability score for each model for both sample size k equaling 5 and 10.
  Results for BETO Spanish model:

```
For k=5 sample size:
Average BETO score per sentence sampled: 0.0215497142
Average mBERT score per sentence sampled: 0.000227659054
```

  Results for MBERT model:

```
For k=10 sample size:
Average BETO score per sentence sampled: 0.0420321291613
Average mBERT score per sentence sampled: 0.0063021180826
```

- Reasonably, the results for k=5 are smaller than k=10 for both models as we are adding a smaller sample of the logits together. However, BETO consistently on average rates randomly sampled tokens from a sentence with a higher probability than mBERT does, not just by a small margin but by two orders of magnitude for k=5 and one order of magnitude for k=10.

6 Analysis

Our results show that BETO has a better grasp of Spanish pro-dropping than mBERT. It is impossible to know for sure if this is purely English influence, but given than English is the at the top for language data quantity and quality, and given previous results from past publications showing that lower resource languages that are typologically distant from English (with different lexicon, morphological structure, word order, etc.) perform worse on benchmarks than English does, it is highly probably that English has influenced mBERT to less readily accept pronoun-dropped sentences to a certain degree.

Pronoun-dropping is a nuanced and intricate process, and the precise conditions vary from language to language. Because BETO has only been exposed to Spanish text, it does not have influences from non pro-drop languages (languages like English that do not allow for pronoun-dropping) or even other pro-drop languages that have different environments. Therefore, just like a monolingual native speaker of a language, even though it may have less generalizeable knowledge that can be

transferred than a non-native polyglot speaker (mBERT), it captures intricacies that come with high exposure to a language. However, just like any non-native speaker, with more exposure to the target language, mBERT can be a powerful tool for language transfer.

7 Conclusion

Trained on over 100 languages, mBERT contains a rich base of language-neutral knowledge that can be transferred to a task, regardless of the specific typological features of the target language. Monolingual models take a long time to train, but they can identify typological intricacies that larger multilingual models have trouble identifying. Pronoun-dropping is just one typological feature that we can explore using mBERT. We plan to continue this research project to see if mBERT produces "English-esque" results in adjective ordering, another very semantically and syntactically complex typological difference between English and Spanish, as well as other features in hopes that this research will help identify the issues with large multilingual models that are less apparent and drive solutions to create better, more nuanced multilingual models.

References

[1] Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In *Association for Computational Linguistics (ACL)*, 2019.

[2] Libovický et al. How language-neutral is multilingual BERT? In *Association for Computational Linguistics (ACL)*, 2019.

[3] Wu et al. Are all languages created equal in multilingual BERT? In *Computation and Language (cs.CL)*, 2019.

[4] Canete et al. Spanish pre-trained bert model and evaluation data. In *PML4DC, ICLR*, 2020.

[5] Virtanen et al. Multilingual is not enough: BERT for Finnish. In *Computation and Language (cs.CL)*, 2019.

[6] Lichtarge et al. Corpora generation for grammatical error correction. In *Association for Computational Linguistics (ACL)*, 2019.