Spanish is one of the top-5 spoken languages.

Not always easy to find Spanish NLP resources.

We provide a Spanish BERT model plus training and evaluation data.
Spanish Pre-trained BERT Model and Evaluation Data
Outline

- Unsupervised pre-training
- Transformers and BERT
- Spanish BERT and Corpus
- Results: Spanish vs Multilingual BERT
Unsupervised Pre-training

● Pick an architecture and tons of text

● Train the architecture with a general LM task (unsupervised)

● Use the weights as a starting point for training with a supervised task
Transformer

[Vaswani et al. 2017]
Weather is nice in Santiago during November.
Weather is nice in Santiago during November.
BERT: Bidirectional Encoder Representations from Transformers

[Delvin et al. 2019]
Weather is nice during November in Santiago.
Weather is nice in Santiago during November.
Unsupervised Pre-training

- **Pick an architecture** and tons of text
- Train the architecture with a **general LM task** (unsupervised)
- Use the weights as a starting point for training with a supervised task
Classification (CLS) token during pre-training

Transformer Encoder

Prediction

Weather is nice during November in Santiago.
Classification token during fine tuning

Positive

Classification

Transformer Encoder

CLS  Weather  is  nice  in  Santiago  during  November
Other BERT technicalities

- **Next Sentence Prediction:**

  ¿Is “Sentence2” the sentence that follows “Sentence1”?

  ![Diagram](CLS Sentence1 SEP Sentence2 SEP)
Other BERT technicalities

● *Positional and Segment Encodings*

● ~15% corruptions during pre-training

● **Base**: 12L - 768H - 12A (110M parameters)

● **Large**: 24L - 1024H - 16A (340M parameters)
Spanish BERT
Spanish Corpus for Pre-training

3 billion words, including:

- Spanish Wikipedia
- Spanish portion of EUBookshop
- Spanish OpenSubtitles, TED Talks, News, etc.

https://github.com/josecannete/spanish-corpora
Spanish BERT

- Whole-Word Masking
- Dynamic Masking
- cased/uncased

https://github.com/dccuchile/beto
Downstream Tasks

- Natural Language Inference
- Paraphrasing
- Named Entity Recognition
- Part of Speech
- Document Classification
- Question Answering

https://github.com/dccuchile/glues
Spanish BERT vs Multilingual BERT
<table>
<thead>
<tr>
<th>Model</th>
<th>XNLI</th>
<th>PAWS-X</th>
<th>NER</th>
<th>POS</th>
<th>MLDoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best mBERT</td>
<td>78.50&lt;sup&gt;a&lt;/sup&gt;</td>
<td>89.00&lt;sup&gt;b&lt;/sup&gt;</td>
<td>87.38&lt;sup&gt;a&lt;/sup&gt;</td>
<td>97.10&lt;sup&gt;a&lt;/sup&gt;</td>
<td>95.70&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>es-BERT uncased</td>
<td>80.15</td>
<td>89.55</td>
<td>82.67</td>
<td>98.44</td>
<td><strong>96.12</strong>&lt;sup&gt;*&lt;/sup&gt;</td>
</tr>
<tr>
<td>es-BERT cased</td>
<td><strong>82.01</strong></td>
<td>89.05</td>
<td><strong>88.43</strong></td>
<td><strong>98.97</strong>&lt;sup&gt;*&lt;/sup&gt;</td>
<td>95.60</td>
</tr>
<tr>
<td>Model</td>
<td>MLQA, MLQA</td>
<td>TAR, XQuAD</td>
<td>TAR, MLQA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------</td>
<td>------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best mBERT</td>
<td>53.90 / 37.40&lt;sup&gt;c&lt;/sup&gt;</td>
<td>77.60 / 61.80&lt;sup&gt;d&lt;/sup&gt;</td>
<td>68.10 / 48.30&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>es-BERT uncased</td>
<td>67.85 / 46.03</td>
<td>77.52 / 55.46</td>
<td>68.04 / 45.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>es-BERT cased</td>
<td><strong>68.01 / 45.88</strong></td>
<td>77.56 / 57.06</td>
<td><strong>69.15 / 45.63</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Future Work

- Add new tasks
- Train smaller models
Spanish Pre-trained BERT Model and Evaluation Data

https://github.com/dccuchile/beto

José Cañete  Gabriel Chaperón  Rodrigo Fuentes  Jou-Hui Ho  Hojing Kang  Jorge Pérez