Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling

Elena Álvarez-Mellado[1] Constantine Lignos[2]

[1]NLP & IR group, UNED [2]Michtom School of Computer Science, Brandeis University

Objective

To detect unassimilated lexical borrowings in Spanish newswire

i.e. words from other languages (mainly English) that have recently been imported into Spanish and that are being used in Spanish newspapers

Ex: *Las prendas bestsellers se estampan con motivos florales, 'animal print' o a retales tipo patchwork**

*Best-seller clothes feature flower print, animal print or patchwork style

What is lexical borrowing?

Lexical borrowing is the incorporation of words form one language into another language.

For ex., using in Spanish words that come from English: *podcast, app, online, crowdfunding, big data, fake news...*

Borrowings in linguistics:

- A source of new words
- Borrowing is a manifestation of how languages change and interact

Borrowings in NLP:

- A source of out-of-vocabulary words
- Automatically detection of lexical borrowings is relevant for NLP downstream tasks: text-to-speech, parsing, machine translation

Set	Tokens	ENG	OTHER	Unique
Training	231,126	1,493	28	380
Development	82,578	306	49	316
Test	58,997	1,239	46	987
Total	372,701	3,038	123	1,683

Table 1: Corpus splits with counts

Annotated corpus

A corpus of Spanish texts annotated with lexical borrowings:

- Composed of a collection of Spanish newspapers
- Annotated with lexical borrowings with 2 tags:
  - `ENG`: for English borrowings
  - `OTHER`: for borrowings from other languages
- BIO encoding: Borrowings can be single token (*app*) or multitoken (*machine learning*)

```
En O
este O
mes O
especialmente O
puede O
ser O
de O
utilidad O
apuntarnos O
al O
batch B-ENG
cooking I-ENG
Benching B-ENG
, O
estar O
en O
el O
banquillo O
de O
tu O
crush B-ENG
mientras O
otro O
juega O
de O
titular O
```

A difficult test set

The main goal of data selection was to create a test set with minimal overlap in words and topics between the training set and the test set, allowing for better assessment of models' generalization. The test set comes from sources and dates not seen in the training set, is very borrowing-dense and contains as many out-of-vocabulary (OOV) words as possible: 92% of the borrowings in the test set are OOV.

Modeling

The corpus was used to evaluate four types of models for borrowing extraction: (1) a CRF model, (2) two Transformer-based models, (3) a BiLSTM-CRF model with different types of unadapted embeddings (word, BPE, and character embeddings) and (4) a BiLSTM-CRF model with previously fine-tuned Transformer-based embeddings pretrained on codeswitched data.

Model	**Word emb.**	**BPE emb.**	**Char emb.**	**Development**			**Test**		
				Prec.	Recall	F1	Prec.	Recall	F1
CRF	`w2v (spa)`	-	-	74.13	59.72	66.15	77.89	43.04	55.44
BETO	-	-	-	73.36	73.46	73.35	86.76	75.50	80.71
mBERT	-	-	-	79.96	73.86	76.76	88.89	76.16	82.02
BiLSTM-CRF	`BETO+BERT`	`en, es`	-	**85.84**	77.07	**81.21**	90.00	76.89	82.92
BiLSTM-CRF	`BETO+BERT`	`en, es`	✓	84.29	**78.06**	81.05	89.71	78.34	83.63
BiLSTM-CRF	`Codeswitch`	-	-	80.21	74.42	77.18	90.05	76.76	82.83
BiLSTM-CRF	`Codeswitch`	-	✓	81.02	74.56	77.62	89.92	77.34	83.13
BiLSTM-CRF	`Codeswitch`	`en, es`	-	83.62	75.91	79.57	90.43	78.55	84.06
BiLSTM-CRF	`Codeswitch`	`en, es`	✓	82.88	75.70	79.10	**90.60**	**78.72**	**84.22**

Table 2: Scores for the development and test sets across all models.

Error analysis

Recall was a weak point for all models. Some of the most frequent false negatives:

- Uppercase borrowings (such as *Big Data*)
- Borrowings in sentence-initial position (in titlecase)
- Words that exist both in English and Spanish (like *primer* or *red*)

Contributions

- A new corpus of Spanish newswire annotated with unassimilated lexical borrowings (more borrowing-dense, OOV-rich)
- Analysis of the perfomance of 4 types of sequence-labeling models trained for lexical borrowing detection:
  - CRF model with handcrafted features
  - Transformer-based models (BETO, mBERT)
  - BiLSTM with Transformer-based word embeddings (BERT+BETO) and subword embeddings (BPE, char)
  - BiLSTM with embeddings pretrained on codeswitched data

More information

- Corpus: `https://github.com/lirondos/coalas`
- HuggingFace models: `https://huggingface.co/models?arxiv=arxiv:2203.16169`
- Paper: `https://arxiv.org/abs/2203.16169`