Extracting English lexical borrowings from Spanish newswire

Elena Álvarez Mellado
USC Information Sciences Institute
890 Winter St, Waltham MA 02451
elena@isi.edu

1 Introduction

Lexical borrowing is a phenomenon that affects all languages and constitutes a productive mechanism for word formation. Previous work on computational detection of lexical borrowings have relied mostly on dictionary and corpora lookup (Alex, 2008; Andersen, 2012; Serigos, 2017), with the limitation that implies. On the other hand, computational approaches to mix-language data have usually framed the task of identifying the language of a word as a tagging problem, where every word in the sequence receives a language tag (Molina et al., 2016; Solorio et al., 2014). In this work we propose to treat lexical borrowing as an extraction problem (in a similar fashion to Named Entity Recognition) in order to build a model that extracts English lexical borrowings from a corpus of Spanish daily news.

In this work, we present: (1) a corpus of European Spanish newswire annotated with anglicisms; (2) a sequence labeling model to extract English lexical borrowings (or *anglicisms*) from Spanish newswire; and (3) a tracking corpus of anglicism usage in the Spanish press.

2 Corpus

A corpus of European Spanish newswire was collected and annotated for the task (Álvarez Mellado, 2020a). The corpus consisted of a collection of monolingual newspaper headlines written in European Spanish. These headlines were extracted from typically anglicism-rich sections: economy, technology, lifestyle, music, TV and opinion.

In addition to the usual train/development/test split, a supplemental test set was collected. The items in the supplemental test did not overlap in time with the main corpus and include more sections. The motivation behind this supplemental test set was to assess the model performance on a more naturalistic setting with a less borrowing-dense sample. The number of tokens and anglicisms per corpus split can be found in Table 1.

Set	Tokens	Anglicisms	Other borrowings
Train	154,632	747	40
Dev	44,758	219	14
Test	44,724	212	13
Suppl. test	81,551	126	35

Table 1: Number of tokens and anglicisms per corpus subset.

The annotation focused on direct, unadapted, emerging anglicisms, i.e. lexical borrowings from the English language into Spanish that have recently been imported and that have still not been assimilated into Spanish (such as *prime time*, *influencer*, *hat-trick*, etc)[1]. Both single-token anglicisms and multiword anglicisms were annotated using the label `ENG`. Additionally, borrowings from other languages other than English were annotated using the label `OTHER`.

3 Model

A sequence labeling model that extracts lexical borrowings from English was trained and tested using the corpus described above. The model chosen was Conditional Random Field (CRF), that was built using `pycrfsuite` (Korobov and Peng, 2014), a Python wrapper for `crfsuite` (Okazaki, 2007) which implements CRF for labeling sequential data. It also used the `Token` and `Span` classes from `spaCy` library (Honnibal and Montani, 2017).

Each word was represented by the following set of binary features (these features are commonly used in NER): bias feature, token feature, uppercase feature, titlecase feature, character trigram

[1] See annotation guidelines in Álvarez Mellado (2020b)

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 384-386.
Held on-line February 14-19, 2021

feature, quotation feature, word suffix feature, POS feature (provided by `spaCy`), word shape feature (provided by `spaCy`) and word2vec Spanish embedding representation (from the Spanish Billion Words Corpus by Cardellino (2019)).

Given that anglicisms can be multiword expressions (such as *best seller, big data*) and that those units should be treated as one borrowing and not as two independent borrowings, multi-token BIO encoding adapted from Ramshaw and Marcus (1999) was used to denote the boundaries of each span. A window of two tokens in each direction was set for the feature extractor. Optimization was performed using Limited-memory BFGS, hyperparameter tuning was done through grid search.

Set	Precision	Recall	F1 score
Development set (− OTHER)	97.84	82.65	89.60
Development set (+ OTHER)			
ENG	96.79	82.65	89.16
OTHER	100.00	28.57	44.44
BORROWING	96.86	79.40	87.26
Test set (− OTHER)	95.05	81.60	87.82
Test set (+ OTHER)			
ENG	95.03	81.13	87.53
OTHER	100.00	46.15	63.16
BORROWING	95.19	79.11	86.41
Supplemental test set (− OTHER)	83.16	62.70	71.49
Supplemental test set (+ OTHER)			
ENG	82.65	64.29	72.32
OTHER	100.00	20.00	33.33
BORROWING	87.62	57.14	69.17

Table 2: Results on development, test set and supplemental test set.

The model produced an F1 score of 89.60 on the development set and 87.82 on test, precision being consistently higher than recall (see Table 2). The results on the supplemental test were significantly lower, with an F1 score of 71.49, which indicates that the difference across topics can have a big impact on the model's performance. These scores were calculated using span level evaluation, which means that only full matches were considered correct and no credit was given to partial matching.

A feature ablation study was done in order to test the contribution of each feature to the model's performance. The ablation study showed that all the handcrafted features contributed to the results, with the character trigram being the one that contributed the most (see Table 3).

The error analysis showed that the model tended to ignore anglicisms appearing in the first position of the sentence, as these words were capitalized and were probably mistaken with proper names. Concerning false positives, the model incorrectly labeled certain neologisms, orthographically adapted borrowings and some proper names as borrowings. English words from film titles and songs were also a common source of mistake. On the other hand, the results also showed that the model was capable of generalising, as it was able to detect lexical borrowings that had never been seen during training.

Features	Precision	Recall	F1 score	F1 change
All features	97.84	**82.65**	**89.60**	
− Bias	96.76	81.74	88.61	−0.99
− Token	95.16	80.82	87.41	−2.19
− Uppercase	97.30	82.19	89.11	−0.49
− Titlecase	96.79	**82.65**	89.16	−0.44
− Char trigram	96.05	77.63	85.86	**−3.74**
− Quotation	97.31	**82.65**	89.38	−0.22
− Suffix	97.30	82.19	89.11	−0.49
− POS tag	**98.35**	81.74	89.28	−0.32
− Word shape	96.79	**82.65**	89.16	−0.44
− Word embedding	95.68	80.82	87.62	−1.98

Table 3: Ablation study results on the development test.

4 A tracking corpus of anglicism usage

The model presented in Section 3 was used to build a continuously-growing corpus that tracks anglicism usage in the daily news of Spain[2].

This tracking corpus consists of newspaper articles from 8 major Spanish newspapers that have been automatically collected on a daily basis since April 2020. The articles are extracted via RSS, preprocessed (for HTML tag removal, etc) and then sent to the CRF model presented in Section 3. The anglicisms extracted by the CRF model are collected and stored in a database. For every anglicism, the date, context, newspaper, and link to the article where the anglicism was found are stored. The database is automatically updated daily and is periodically revised by a human to remove and correct errors. At the time of writing this document, the database stores more than 110,000 borrowings (8,000 distinct borrowings), collected since April 2020. The database can be queried through the project's website and can also be downloaded as CSV files.

This automatically collected corpus can inform language change by monitoring anglicism usage and detecting novel anglicisms that appear in the Spanish press.

5 Conclusions

We have presented a novel approach to lexical borrowing detection that frames the problem as an

[2]See `http://observatoriolazaro.es/en/`

extraction task (rather than as a tagging task) and we have applied it to extract English lexical borrowings from a new annotated corpus of Spanish newswire. The proposed model is a CRF model that uses handcrafted features (similar to those used in NER models). Unlike prior work, the model we have introduced doesn't rely on lexicon or corpus lookup. The results show that this is a productive approach to borrowing extraction, that can also sucessfully extract previously unseen anglicisms as well as multiword lexical borrowings.

Finally, the automatically collected corpus can inform language change and help us understand more about language contact in general and the process of borrowing in particular.

References

Beatrice Alex. 2008. *Automatic detection of English inclusions in mixed-lingual data with an application to parsing*. Ph.D. thesis, University of Edinburgh.

Elena Álvarez Mellado. 2020a. An Annotated Corpus of Emerging Anglicisms in Spanish Newspaper Headlines. In *Proceedings of the Fourth Workshop on Computational Approaches to Code Switching*, pages 1–8, Marseille, France. European Language Resources Association.

Elena Álvarez Mellado. 2020b. Lázaro: An extractor of emergent anglicisms in Spanish newswire. Master's thesis, Brandeis University.

Gisle Andersen. 2012. Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Cristiano Furiassi, Virginia Pulcini, and Félix Rodríguez González, editors, *The anglicization of European lexis*, pages 111–130.

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings. https://crscardellino.github.io/SBWCE/.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/.

M Korobov and T Peng. 2014. Python-crfsuite. https://github.com/scrapinghub/python-crfsuite.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Jacqueline Rae Larsen Serigos. 2017. *Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish*. Ph.D. thesis, The University of Texas at Austin.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.