Extracting English lexical borrowings from Spanish newswire

Elena Álvarez-Mellado[1]

[1]USC Information Sciences Institute

Objectives

Build a model that can extract English lexical borrowings (or *anglicisms*) from a corpus of Spanish daily news.

For that we have developed:

❶ A corpus of Spanish newswire annotated with anglicisms.

❷ A sequence labeling model that can extract English. lexical borrowings.

❸ A continuously-growing corpus that tracks anglicism usage in the daily news of Spain.

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 framed the task as a tagging problem (where each word receives a tag) and relied on dictionary and corpora lookup [1, 2, 3], with the limitation that implies.

We propose to treat lexical borrowing as an extraction problem (in a similar fashion to Named Entity Recognition).

Corpus

A corpus of Spanish newswire was collected and annotated [4].

- non-assimilated anglicisms
- single-token and multitoken
- example: *prime time*, *influencer*, *hat-trick*

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.

Model

The corpus was used to train a CRF model with handcrafted features (see Table 2) that extracts English lexical borrowings.

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 2:Ablation study results on the development test.

Application: A tracking corpus of anglicism usage

The CRF model was used to build a continuously-growing corpus that tracks anglicism usage in the daily news of Spain (see `http://observatoriolazaro.es/en/`).

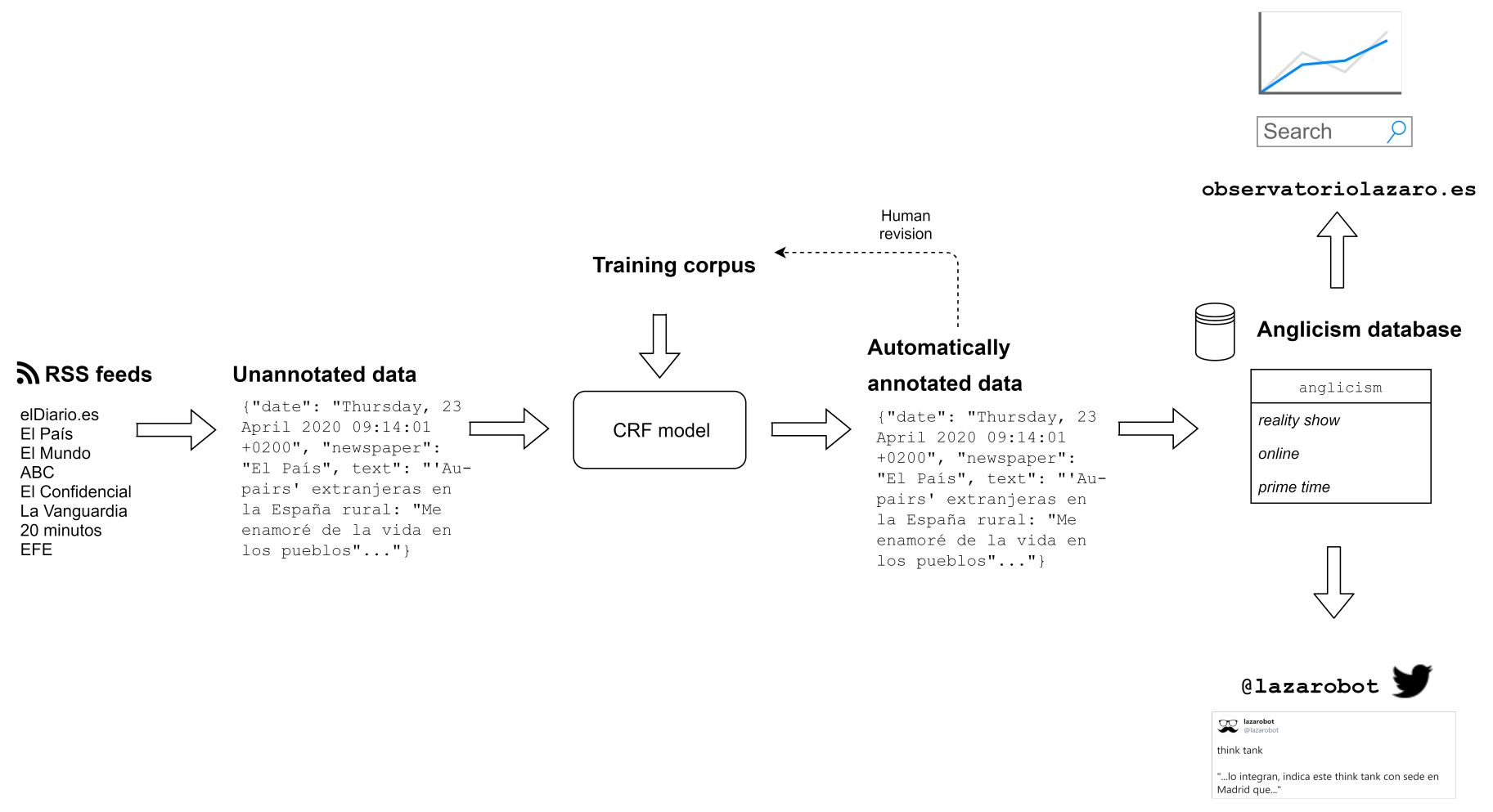
RSS feeds
elDiario.es
El País
El Mundo
ABC
El Confidencial
La Vanguardia
20 minutos
EFE
Unannotated data
{"date": "Thursday, 23 April 2020 09:14:01 +0200", "newspaper": "El País", text": "'Au-pairs' extranjeras en la España rural: "Me enamoré de la vida en los pueblos"..."}
Training corpus
CRF model
Human revision
Automatically annotated data
{"date": "Thursday, 23 April 2020 09:14:01 +0200", "newspaper": "El País", text": "'Au-pairs' extranjeras en la España rural: "Me enamoré de la vida en los pueblos"..."}
Anglicism database
anglicism
reality show
online
prime time
observatoriolazaro.es
Search
@lazarobot
lazarobot
@lazarobot
think tank
"...lo integran, indica este think tank con sede en Madrid que..."

Figure 1:Automatic pipeline of anglicism extraction.

Lexical borrowing detection as an extraction task

We propose to approach lexical borrowing detection as an extraction task (*à la NER*), instead of as a tagging problem (*à la POS-tagging*) in order to build a model that can extract novel English lexical borrowings (both single-token and multi-token) from a corpus of Spanish newswire.

Model results

Results obtained on the different sets of the corpus:

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 3:Results on dev set, test set and supplemental test set.

Extraction pipeline

- 8 major Spanish newspapers are automatically scraped daily since April 2020.
- The articles are extracted via RSS, preprocessed (for HTML tag removal, etc) and then sent to the CRF model.
- The anglicisms extracted by the CRF model are collected and stored in a database.
- For every anglicism, 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

Lexical database & Visualizations

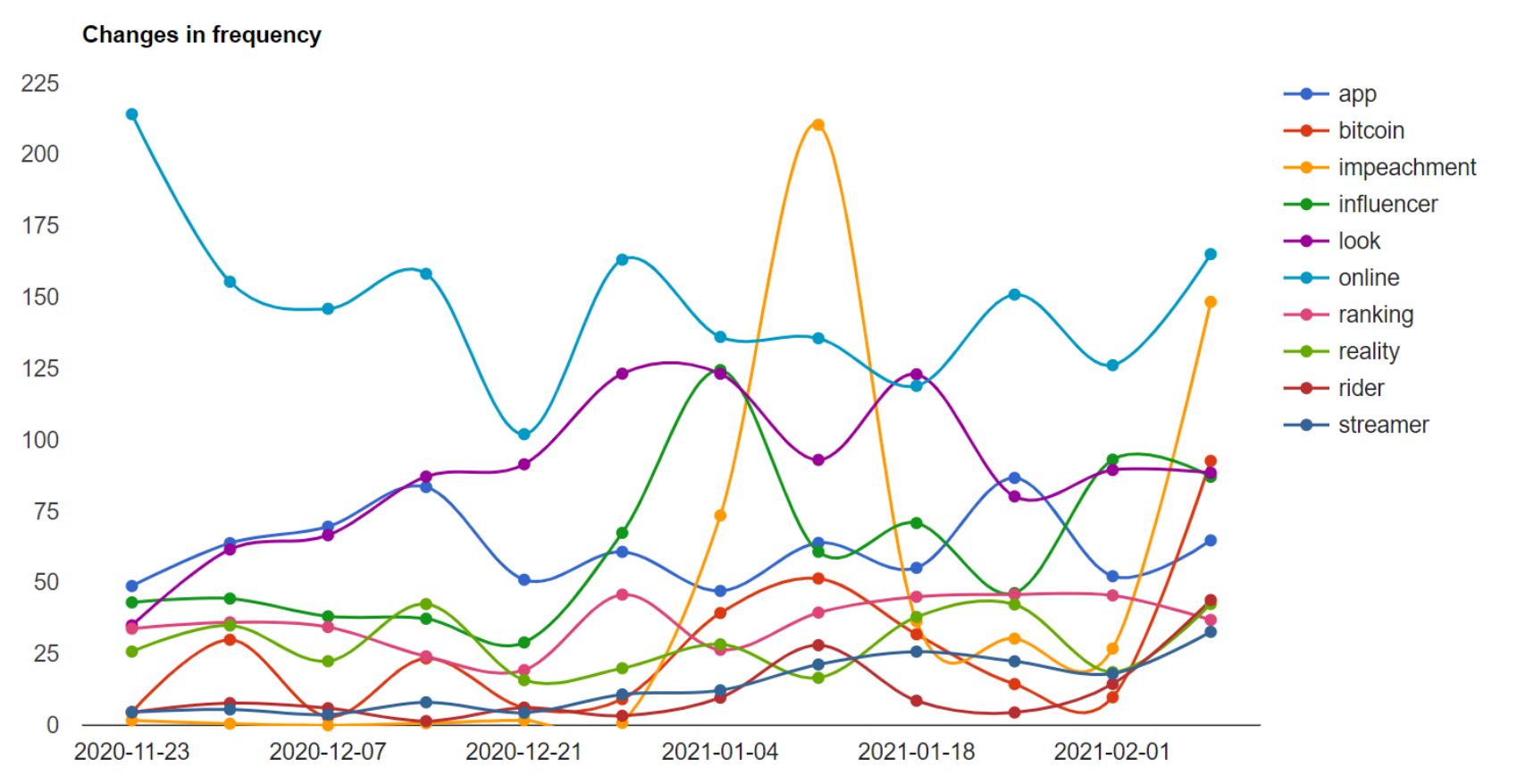

Figure 2:Changes in frequency of the most frequent anglicisms

Conclusions

- Borrowing extraction can be framed as an extraction problem (*à la NER*).
- We train a CRF model with handcrafted features to extract English lexical borrowings from a corpus of Spanish newswire.
- The model doesn't rely on lexicon or corpus lookup.
- The model can extract previously unseen anglicisms and multiword lexical borrowings.

References

[1] Beatrice Alex.
Automatic detection of English inclusions in mixed-lingual data with an application to parsing.
PhD thesis, University of Edinburgh, 2008.

[2] Gisle Andersen.
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. 2012.

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

[4] Elena Álvarez Mellado.
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, May 2020. European Language Resources Association.

More information

- `https://observatoriolazaro.es/en`
- `elena@isi.edu`
- `@lazarobot`