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

ACL 2022

Table of Contents

Table of Contents

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, spin-off, big data, fake news...

- Lexical borrowing is a type of linguistic borrowing.
  - Linguistic borrowing is the process of reproducing in one language the patterns of other languages Haugen (1950)
- Borrowing and code-switching are related and have frequently been described as a continuum Clyne et al. (2003)
  - Code-switching = mixing two languages in one sentence.
    Ex: *You start a sentence in English y la acabas en español*
    Poplack (1980); Poplack et al. (1988)

Lexical borrowing vs Code switching

	Code Switching	**Lexical Borrowing**
Speaker	bilinguals	monolinguals
Grammar compliance	both languages	recipient language
Level of integration	not integrated	can be integrated
NLP approach	one tag per token (*à la POS-tagging*)[1]	extraction of spans of interest (*à la NER*)

[1] see Computational Approaches to Linguistic Code-Switching workshops (CALCS) Solorio et al. (2014); Diab et al. (2016); Aguilar et al. (2018); Solorio et al. (2020, 2021)

Why is borrowing an interesting phenomenon?

Borrowing in Linguistics:

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)

Why is borrowing an interesting phenomenon?

Borrowing in Linguistics:

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)
- Lexical borrowings are a source of new words

  online, software, streaming...

Why is borrowing an interesting phenomenon?

Borrowing in Linguistics:

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)
- Lexical borrowings are a source of new words

  online, software, streaming...

Borrowing in NLP:

- Borrowings are a common source of out-of-vocabulary words Gerding Salas et al. (2018).

Why is borrowing an interesting phenomenon?

Borrowing in Linguistics:

- Borrowing is a manifestation of how languages change and interact (diachronic linguistics and contact linguistics) Weinreich (1963)
- Lexical borrowings are a source of new words

  online, software, streaming...

Borrowing in NLP:

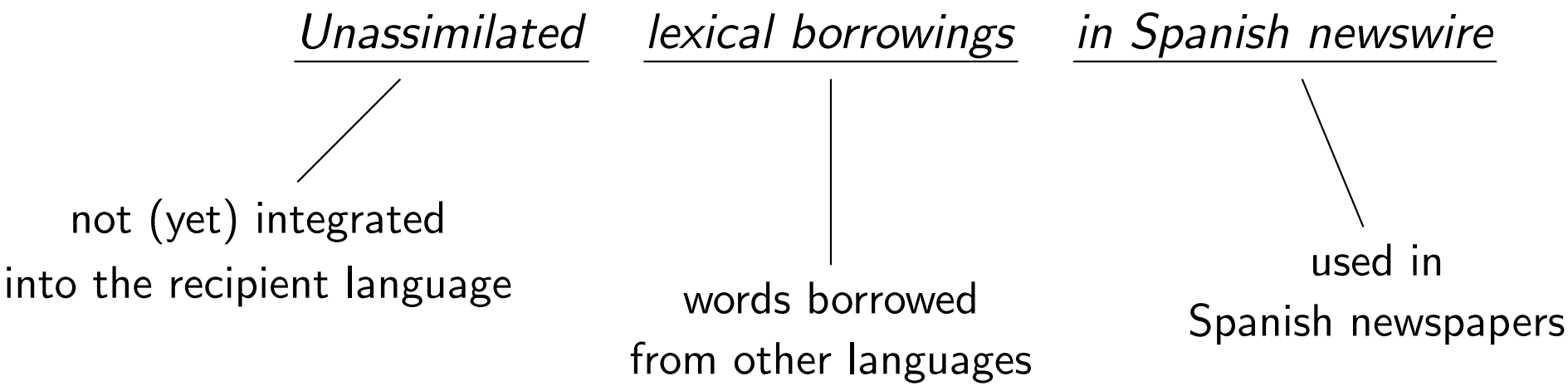
- Borrowings are a common source of out-of-vocabulary words Gerding Salas et al. (2018).
- Automatically detecting lexical borrowings from text has proven to be relevant for NLP downstream tasks:
  - Parsing Alex (2008)
  - Text-to-speech synthesis Leidig et al. (2014)
  - Machine translation Tsvetkov and Dyer (2016)

Previous work on English borrowing detection

- Growing interest in the influence of English in other languages Görlach (2002).
- With a focus on English lexical borrowings (a.k.a *anglicisms*)
- Previous work on automatic detection of borrowings in different European languages: German, French, Italian, Norwegian, Finnish, Spanish Andersen (2012); Chesley (2010); Furiassi and Hofland (2007); Garley and Hockenmaier (2012); Losnegaard and Lyse (2012); Mansikkaniemi and Kurimo (2012); Serigos (2017); Álvarez Mellado (2020)

Table of Contents

The task



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 show flower print, animal print or patchwork style

Limitations in previous work on anglicism detection

Previous work introduced a CRF model for anglicism detection on Spanish newswire (F1=86) (Álvarez Mellado, 2020). However, both the dataset and modeling approach had significant limitations:

- The dataset consisted only of headlines.
- The dataset focused exclusively on a single source of news
- The number and variety of borrowings were limited
- There was a significant overlap in borrowings between the training set and the test set (which prevented assessment of whether the modeling approach was actually capable of generalizing to previously unseen borrowings)
- The best results were obtained by a CRF model, and more sophisticated approaches were not explored.

Table of Contents

Dataset creation

The aim was to create a test set as difficult as possible with minimal overlap in words and topics between the training set and the test set.

The test set:

- comes from sources and dates not seen in the training set

Dataset creation

The aim was to create a test set as difficult as possible with minimal overlap in words and topics between the training set and the test set.

The test set:

- comes from sources and dates not seen in the training set

  Training set: Aug-Dec 2020; Test set: Feb-March 2021

Dataset creation

The aim was to create a test set as difficult as possible with minimal overlap in words and topics between the training set and the test set.

The test set:

- comes from sources and dates not seen in the training set

  Training set: Aug-Dec 2020; Test set: Feb-March 2021
- is very borrowing-dense

Dataset creation

The aim was to create a test set as difficult as possible with minimal overlap in words and topics between the training set and the test set.

The test set:

- comes from sources and dates not seen in the training set

  Training set: Aug-Dec 2020; Test set: Feb-March 2021

- is very borrowing-dense

  Training set: 6 bor/1,000 tokens. Test set: 20 bor/1,000 tokens

Dataset creation

The aim was to create a test set as difficult as possible with minimal overlap in words and topics between the training set and the test set.

The test set:

- comes from sources and dates not seen in the training set

  Training set: Aug-Dec 2020; Test set: Feb-March 2021
- is very borrowing-dense

  Training set: 6 bor/1,000 tokens. Test set: 20 bor/1,000 tokens
- contain as many out-of-vocabulary (OOV) words as possible

Dataset creation

The aim was to create a test set as difficult as possible with minimal overlap in words and topics between the training set and the test set.

The test set:

- comes from sources and dates not seen in the training set

  Training set: Aug-Dec 2020; Test set: Feb-March 2021
- is very borrowing-dense

  Training set: 6 bor/1,000 tokens. Test set: 20 bor/1,000 tokens
- contain as many out-of-vocabulary (OOV) words as possible

  92% of the borrowings in the test set are OOV

The corpus

The corpus was:

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

The corpus: counts

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: Corpus splits with counts

The corpus: example

```
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
```

Table of Contents

Conditional Random Fields model

A CRF model with handcrafted features. The following set of binary features from (Álvarez Mellado, 2020) were used:

– Bias
– Token
– Uppercase
– Titlecase
– Character trigram
– Quotation marks
– Suffix
– POS tag (provided by `spaCy`)
– Word shape (provided by `spaCy`)
– Word embedding (Spanish word2vec by Cardellino (2019))
– URL (provided by `spaCy`)
– Email (provided by `spaCy`)
– Twitter (`#hashtag` or `@username`)

CRF results

Previous work with a similar CRF on a different dataset had reported and F1 score of 86.41 (Álvarez Mellado, 2020). We got F1=55.44.

Set	Precision	Recall	F1
Development			
ALL	74.13	59.72	66.15
ENG	74.20	68.63	71.31
OTHER	66.67	4.08	7.69
Test			
ALL	77.89	43.04	55.44
ENG	78.09	44.31	56.54
OTHER	57.14	8.70	15.09

Table: CRF performance on the development and test sets

Transformer-based models

We evaluated two Transformer-based models:

- BETO base cased model: a monolingual BERT model trained for Spanish (Cañete et al., 2020)
- mBERT: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2019)

Both models were run using the `Transformers` library by HuggingFace (Wolf et al., 2020).

Transformer-based models results

	Development			Test		
	Precision	Recall	F1	Precision	Recall	F1
BETO						
ALL	73.36	73.46	73.35	86.76	75.50	80.71
ENG	74.30	84.05	78.81	87.33	77.99	82.36
OTHER	47.24	7.34	11.93	36.12	8.48	13.23
mBERT						
ALL	**79.96**	**73.86**	**76.76**	**88.89**	**76.16**	**82.02**
ENG	80.25	84.31	82.21	89.25	78.85	83.64
OTHER	66.18	8.6	14.41	45.30	7.61	12.84

BiLSTM-CRF model with different types of embeddings

Can a BiLSTM-CRF model fed with different embeddings that encode different linguistic information outperform the results obtained by the Transformer-based models?

BiLSTM-CRF model with different types of embeddings

Can a BiLSTM-CRF model fed with different embeddings that encode different linguistic information outperform the results obtained by the Transformer-based models?

We ran some preliminary experiments with different types of embeddings: Transformer-based, FastText, one-hot, byte pair and character embeddings.

BiLSTM-CRF model with different types of embeddings

Can a BiLSTM-CRF model fed with different embeddings that encode different linguistic information outperform the results obtained by the Transformer-based models?

We ran some preliminary experiments with different types of embeddings: Transformer-based, FastText, one-hot, byte pair and character embeddings.

This is what we found out:

- Transformer-based embeddings $>$ non-contextualized embeddings

BiLSTM-CRF model with different types of embeddings

Can a BiLSTM-CRF model fed with different embeddings that encode different linguistic information outperform the results obtained by the Transformer-based models?

We ran some preliminary experiments with different types of embeddings: Transformer-based, FastText, one-hot, byte pair and character embeddings.

This is what we found out:

- Transformer-based embeddings $>$ non-contextualized embeddings
- English BERT $+$ Spanish BETO embeddings $>$ mBERT embeddings

BiLSTM-CRF model with different types of embeddings

Can a BiLSTM-CRF model fed with different embeddings that encode different linguistic information outperform the results obtained by the Transformer-based models?

We ran some preliminary experiments with different types of embeddings: Transformer-based, FastText, one-hot, byte pair and character embeddings.

This is what we found out:

- Transformer-based embeddings $>$ non-contextualized embeddings
- English BERT $+$ Spanish BETO embeddings $>$ mBERT embeddings
- BPE embeddings $\Longrightarrow$ better F1

BiLSTM-CRF model with different types of embeddings

Can a BiLSTM-CRF model fed with different embeddings that encode different linguistic information outperform the results obtained by the Transformer-based models?

We ran some preliminary experiments with different types of embeddings: Transformer-based, FastText, one-hot, byte pair and character embeddings.

This is what we found out:

- Transformer-based embeddings $>$ non-contextualized embeddings
- English BERT $+$ Spanish BETO embeddings $>$ mBERT embeddings
- BPE embeddings $\Longrightarrow$ better F1
- Character embeddings $\Longrightarrow$ better recall

Best BiLSTM-CRF results

Embeddings	**Development**			**Test**		
	Precision	Recall	F1	Precision	Recall	F1
BETO+BERT and BPE						
ALL	**85.84**	77.07	**81.21**	**90.00**	76.89	82.92
ENG	86.15	88.00	87.05	90.20	79.36	84.42
OTHER	72.81	8.8	15.60	62.68	10.43	17.83
BETO+BERT, BPE, and char						
ALL	84.29	**78.06**	81.05	89.71	**78.34**	**83.63**
ENG	84.54	89.05	86.73	89.90	80.88	85.14
OTHER	73.50	9.38	16.44	61.14	9.78	16.81

(Best results with mBERT obtained F1=76 on the dev set and F1=82 on the test set. So yes, it seems that a BiLSTM-CRF fed with different embeddings could outperform mBERT)

BiLSTM-CRF model with codeswitch embeddings

Can lexical borrowing detection be framed as transfer learning from language identification in codeswitching?

BiLSTM-CRF model with codeswitch embeddings

Can lexical borrowing detection be framed as transfer learning from language identification in codeswitching?

We ran a BiLSTM-CRF model but instead of using the unadapted Transformer embeddings, we used codeswitch embeddings.

BiLSTM-CRF model with codeswitch embeddings

Can lexical borrowing detection be framed as transfer learning from language identification in codeswitching?

We ran a BiLSTM-CRF model but instead of using the unadapted Transformer embeddings, we used codeswitch embeddings.

Codeswitch embeddings (Sarker, 2020) = Fine-tuned Transformer-based embeddings pretrained for language identification on the Spanish-English section of the LinCE codeswitching dataset (Aguilar et al., 2020)

BiLSTM-CRF with codeswitching embeddings

Embeddings	**Development**			**Test**		
	Precision	Recall	F1	Precision	Recall	F1
Codeswitch						
`ALL`	80.21	74.42	77.18	90.05	76.76	82.83
`ENG`	80.19	85.59	82.78	90.05	79.37	84.33
`OTHER`	85.83	4.70	8.78	90.00	6.52	12.14
Codeswitch + char						
`ALL`	81.02	74.56	77.62	89.92	77.34	83.13
`ENG`	81.00	85.91	83.34	89.95	80.00	84.67
`OTHER`	73.00	3.67	6.91	68.50	5.43	9.97
Codeswitch + BPE						
`ALL`	**83.62**	**75.91**	**79.57**	90.43	78.55	84.06
`ENG`	83.54	86.86	85.16	90.57	81.14	85.59
`OTHER`	94.28	7.55	13.84	67.17	8.70	15.30
Codeswitch + BPE + char						
`ALL`	82.88	75.70	79.10	**90.60**	**78.72**	**84.22**
`ENG`	82.90	86.57	84.66	90.76	81.32	85.76
`OTHER`	87.23	7.75	14.03	66.50	8.70	15.13

(Best results with the BiLSTM-CRF fed with unadapted embeddings obtained F1=81 on the dev set and F1=83 on the test set)

Table of Contents

Discussion and wrap-up

- A new dataset of Spanish newswire annotated with unassimilated lexical borrowings (more borrowing-dense, OOV-rich)
- 4 types of models for lexical borrowing detection:
  - CRF model with handcrafted features (F1=55)
  - Transformer-based models (BETO: F1=80, mBERT: F1=82)
  - BiLSTM with Transformer-based word embeddings (BERT+BETO) and subword embeddings (BPE, char) (F1=83.6)
  - BiLSTM with embeddings pretrained on codeswitched data (F1=84.2)
- Error analysis: Recall was a weak point for all models. Most frequent false negatives:
  - upper-case borrowings (such as *Big Data*)
  - borrowings in sentence-initial position (in titlecase)
  - words that exist both in English and Spanish (like *primer* or *red*)
- BPE embeddings seem to improve F1 score.
- Character embeddings seem to improve recall

Resources

- Corpus

  `https://github.com/lirondos/coalas`

- HuggingFace models

  `https://huggingface.co/models?arxiv=arxiv:2203.16169`

- Paper

  `https://arxiv.org/abs/2203.16169`

References

Aguilar, G., AlGhamdi, F., Soto, V., Diab, M., Hirschberg, J., and Solorio, T. (2018). Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Aguilar, G., Kar, S., and Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1803–1813, Marseille, France. European Language Resources Association.

Alex, B. (2008). Comparing corpus-based to web-based lookup techniques for automatic English inclusion detection. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Álvarez Mellado, E. (2020). Lázaro: An extractor of emergent anglicisms in Spanish newswire.

Andersen, G. (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Furiassi, C., Pulcini, V., and Rodríguez González, F., editors, The anglicization of European lexis, pages 111–130.

Cardellino, C. (2019). Spanish Billion Words Corpus and Embeddings. `https://crscardellino.github.io/SBWCE/`.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. In PML4DC at ICLR 2020.

Chesley, P. (2010). Lexical borrowings in French: Anglicisms as a separate phenomenon. Journal of French Language Studies, 20(3):231–251.

Clyne, M., Clyne, M. G., and Michael, C. (2003). Dynamics of language contact: English and immigrant languages. Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diab, M., Fung, P., Ghoneim, M., Hirschberg, J., and Solorio, T., editors (2016). Proceedings of the Second Workshop on Computational Approaches to Code Switching, Austin, Texas. Association for Computational Linguistics.

Furiassi, C. and Hofland, K. (2007). The retrieval of false anglicisms in newspaper texts. In Corpus Linguistics 25 Years On, pages 347–363. Brill Rodopi.

References (cont.)

Garley, M. and Hockenmaier, J. (2012). Beefmoves: Dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 135–139, Jeju Island, Korea. Association for Computational Linguistics.

Gerding Salas, C., Cañete González, P., and Adam, C. (2018). Neología sintagmática anglicada en español: Calcos y préstamos. Revista signos, 51(97):175–192.

Görlach, M. (2002). English in Europe. OUP Oxford.

Haugen, E. (1950). The analysis of linguistic borrowing. Language, 26(2):210–231.

Leidig, S., Schlippe, T., and Schultz, T. (2014). Automatic detection of anglicisms for the pronunciation dictionary generation: a case study on our German IT corpus. In Spoken Language Technologies for Under-Resourced Languages.

Losnegaard, G. S. and Lyse, G. I. (2012). A data-driven approach to anglicism identification in Norwegian. In Andersen, G., editor, Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian, pages 131–154. John Benjamins Publishing.

Mansikkaniemi, A. and Kurimo, M. (2012). Unsupervised vocabulary adaptation for morph-based language models. In Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pages 37–40. Association for Computational Linguistics.

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics, 26(1):47–104.

Sarker, S. (2020). Codeswitch. `https://github.com/sagorbrur/codeswitch`.

Serigos, J. R. L. (2017). Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish. PhD thesis, The University of Texas at Austin.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (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.

References (cont.)

Solorio, T., Chen, S., Black, A. W., Diab, M., Sitaram, S., Soto, V., Yilmaz, E., and Srinivasan, A., editors (2021). Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, Online. Association for Computational Linguistics.

Solorio, T., Choudhury, M., Bali, K., Sitaram, S., Das, A., and Diab, M., editors (2020). Proceedings of the The 4th Workshop on Computational Approaches to Code Switching, Marseille, France. European Language Resources Association.

Tsvetkov, Y. and Dyer, C. (2016). Cross-lingual bridges with models of lexical borrowing. Journal of Artificial Intelligence Research, 55:63–93.

Weinreich, U. (1963). Languages in contact (1953). The Hague: Mouton.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.