BERT4EVER at ADoBo 2021: Detection of Borrowings in the Spanish Language Using Pseudo-label Technology

Shengyi Jiang[1,2], Tong Cui[1], Yingwen Fu[1], Nankai Lin[1(✉)] and Jieyi Xiang[1]

[1] School of Information Science and Technology, Guangdong University of Foreign Studies, China
[2] Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou
neakail@outlook.com

Abstract. In this paper, we report the solution of the team BERT 4EVER for the automatic detection of borrowings in the Spanish Language task in IberLeF 2021, which aims to detect lexical borrowings that appear in the Spanish press. We adopt the CRF model to tackle the problem. In addition, we introduce pseudo-label technology and ensemble learning to improve the generalization capability. Experimental results demonstrate the effectiveness of CRF model and pseudo-label technology.

Keywords: Automatic Detection of Borrowings, CRF, Pseudo-label Technology.

1 Introduction

Lexical borrowing is a word formation that is widely used in many languages. Previous work on computational detection of lexical borrowings has relied mostly on dictionary and corpora lookup [1][2][3], with the limitation coming from the original dictionary or corpora. On the other hand, computational approaches to mixed-language data have usually framed the task of identifying the language of a word as a sequence labeling task, where every word in the sequence is attached to a language tag [4][5].

IberLeF 2021 proposes the task "Automatic Detection of Borrowings in the Spanish Language" [6]. Our team, BERT 4EVER, also participates in this task. In this report, we will review our solution to this task, namely, the CRF model aided by pseudo-label technology and ensemble learning.

IberLEF 2021, September 2021, Málaga, Spain.

2 Related Work

Linguistic borrowing is the process of copying elements and patterns from another language into one [7]. This classification system is based on two processes: import and substitution. Input is the incorporation into the recipient's language of a foreign form that may or may not contain a meaning. Substitution refers to the substitution of foreign phonemes or morphemes by foreign phonemes or morphemes of the recipient language so as to localize the foreign form. Both processes can occur in the same borrowings. Thus, linguistic borrowing involves communication between two languages and has been extensively studied in the field of contact linguistics [8]. Various typologies have been proposed to classify language loanwords according to different criteria, such as typological features, linguistic hierarchy involved, integration of loanwords elements in the recipient's language, etc [9][10][11].

Now that English has established itself as the global lingua franca, many languages are currently undergoing the process of importing new loanwords from English. In the past decade, English has produced a large number of lexical loanwords in many European languages, especially in the press.

Previous work on computational detection of lexical borrowings have relied mostly on dictionary and corpora lookup. Studies on anglicization have begun to use a multi-million-word corpus [12][13][14]. Alex [16] combined lexicon lookup and a search engine module that used the web as a corpus to detect English inclusions in a German text corpus and compared the proposed model with a maxent Markov model. Furiassi and Hofland [17] explored corpora lookup and character n-grams to extract false anglicisms from an Italian newspaper corpus. Andersen [2] used dictionary lookup, regular expressions and lexicon-derived frequencies of character n-grams to detect anglicism candidates in the Norwegian Newspaper Corpus (NNC). In computational approaches to mix-language data, the task, aiming for the identification of the language of a word, has usually been assumed as a tagging problem which needs every word in the sequence to be tagged [5].

The large amount of available data presents methodological challenges to data processing for English language research. Corpus-based studies of English borrowings in Spanish media have traditionally relied on manual evaluation of either previously compiled general corpora such as CREA [15], or new tailor-made corpora designed to analyze specific genres, varieties or phenomena. In Spanish, Serigos [18] extracted anglicisms from an Argentinian newspaper corpus by combining dictionary lookup (aided by TreeTagger and the NLTK lemmatizer) with automatic filtering of capitalized words and manual inspection. In Serigos [3], a character n-gram module was added in the dictionary lookup method to estimate the probabilities of a word being English or Spanish. Moreno Fernandez and Moreno Sandoval[19] used different pattern-matching filters and lexicon lookup to extract anglicism candidates from a tweet corpus in US Spanish.

3 Method

In the automatic detection of borrowings in the Spanish Language task, we train five CRF models based on the five-fold data and then use the trained CRF models to predict unlabeled samples. We gather the pseudo-labeled dataset together with the training set to train the new CRF model.

3.1 CRF

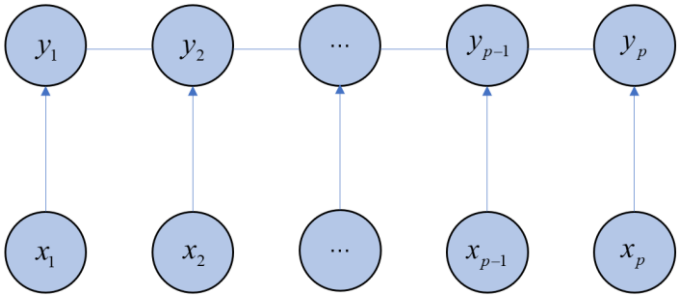

Fig. 1. CRF Model.

There are two random variables, X is a random variable on the sequence of data to be labeled, and Y is a random variable on the corresponding sequence of data to be labeled. The random variables X and Y are under the common distribution, but we construct a conditional model $p(Y|X)$ from paired observation and label sequences in a discriminative framework, without explicitly modeling the marginal $p(X)$.

Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G.Then (X, Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph:

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$$

where $w \sim v$ means that w and v are neighbors in G, $w \neq v$ means all vertices except v. Y_v and Y_w are random variables corresponding to v and w.

Table 1 lists the feature set which obtained the best performance in our experiments, and we report the experiment result based on this feature set. The "collocation" feature is the Co-occurrence between the current word and the next (or last) word.

Table 1. The defined feature sets used in CRF.

Type	Feature	Description
Unigram	$w_n(n = -1,0,1)$	The previous n, current, and next n words
Prefix	$p_n(w_0), n = 2,3,4$	The first n letters in the current word
Suffix	$s_n(w_0), n = 2,3,4$	The last n letters in the current word
Collocation	$w_{n-1}w_n(n = 0,1)$	The collocation of the current word and the next (or last) word

3.2 Ensemble learning

We conduct five-fold cross-validation for the training data and then train five models based on the five-fold data. Each model predicts the test data separately. For each token x, the predicted output of the model is

$$y_i = CRF_i(z)$$

in which z is the token x's feature representation, CRF_i is the i-th CRF model and y_i is the output of i-th CRF model. Therefore, the output of the five models is

$$Y = [y_1, y_2, y_3, y_4, y_5]$$

We consider the label that appears most in Y as the label of x.

3.3 Pseudo-label technology

We use a pseudo-label strategy [20][21] to generate labeled data that does not require manual labeling, as shown in Figure 2. We first use the competition open training set to train CRF models, and then use the trained CRF models to predict unlabeled samples, the predicted results as the sample label. And then we screen all the predicted samples to filter out the sentences without lexical borrowings, only the sentences with lexical borrowings exist. The unlabeled samples we use from GlobalVoices (Spanish portion of GlobalVoices)[2] and News-Commentary11 (Spanish portion of NCv11)[3]. We gather the filtered sentence set together with the training set to train the new CRF model, and although the sample quality obtained through data enhancement is not high, the new model has higher generalization capability to some extent because the new model trained with more data.

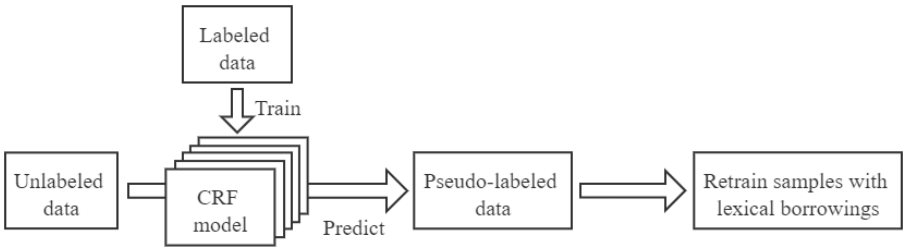

Fig. 2. Pseudo-label Technology Flow Chart.

4 Results

[2] http://opus.nlpl.eu/GlobalVoices.php
[3] http://opus.nlpl.eu/News-Commentary-v11.php

Table 2. The experiment of exploring different Prefix/Suffix performance.

Prefix/ Suffix	P	R	F
/	**91.21%**	8.11%	14.90%
2	73.06%	40.73%	52.29%
3	71.27%	30.76%	42.95%
4	83.15%	20.68%	33.08%
All	76.33%	**41.13%**	53.42%

We first explore the performance of different prefixes/suffixes, and the results are shown in Table 2. The feature of first (and last) 2 letters in the current word has the greatest impact on the task, with an increase of 37.39 in the F value. When all the prefix/suffix features are used together, the effect is the best, and the result based on five-fold cross-validation has reached 53.42%.

Table 3. The results of our model based on five-fold cross-validation.

Model	P	R	F
CRF	**76.33%**	41.13%	53.42%
CRF + Pseudo-label Technology	67.82%	**42.25%**	52.06%

Table 4. The results of our model on final test set.

Model	Type	P	R	F
CRF	ENG	76.48%	25.99%	38.80%
	OTHER	60.00%	6.52%	11.76%
	ALL	76.29%	25.29%	37.99%
CRF + Data Augmentation	ENG	75.43%	28.25%	41.10%
	OTHER	60.00%	6.52%	11.76%
	ALL	75.27%	27.47%	40.25%

As shown in Table 3 and Table 4, the recall of CRF based on pseudo-label technology is significantly improved, which proves that the pseudo-label technology can improve the generalization performance of the model. On the final test set, the F value of the CRF model reached 37.99%, and the F value of the CRF model based on pseudo-label technology is 40.25%, which shows that the pseudo-label technology has a significant impact on detection of borrowings in the Spanish language task.

5 Conclusion

In the automatic detection of borrowings in the Spanish Language task in IberLeF 2021, we adopt the CRF model aided by pseudo-label technology and ensemble learning. In addition, we also explore the impact of different features on the task. In the future, we will try to combine pseudo-label technology with deep learning models in order to achieve better results on the detection of borrowings tasks.

Acknowledgements

This work was supported by the Key Field Project for Universities of Guangdong Province (No. 2019KZDZX1016), the National Natural Science Foundation of China (No. 61572145) and the National Social Science Foundation of China (No. 17CTQ045). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Alex, B.: Automatic detection of English inclusions in mixed-lingual data with an application to parsing. University of Edinburgh (2008).
2. Andersen, G.: Semi-automatic approaches to Anglicism detection in Norwegian corpus data. The anglicization of European lexis, 111-130 (2012).
3. Serigos, J. R. L.: Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish. University of Texas at Austin (2017).
4. Molina, G., AlGhamdi, F., Ghoneim, M., et al.: 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, 40–49 (2019).
5. Solorio, T., Blair, E., Maharjan, S., et al.: 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, pp. 62–72. (2014).
6. Alvarez Mellado, E., Espinosa Anke, L., Gonzalo Arroyo, J., Lignos, C., and Porta Zamorano, J.: Overview of ADoBo 2021 shared task: Automatic Detection of Unassimilated Borrowings in the Spanish Press. Procesamiento del Lenguaje Natural, 67 (2021).
7. Haugen, E.: The analysis of linguistic borrowing. Language 26(2), 210–231 (1950).
8. Weinreich, U.: Languages in contact. Findings and Problems (1953).
9. Haspelmath, M. and Tadmor, U.: Loanwords in the world's languages: a comparative handbook. Walter de Gruyter (2009).
10. Matras, Y. and Sakel, J.: Grammatical borrowing in cross-linguistic perspective.Walter de Gruyter 38 (2007).
11. Thomason, S. G. and Kaufman, T.: Language contact, creolization, and genetic linguistics. Univ of California Press (1992).
12. Andersen, G.: Pragmatic borrowing. Journal of Pragmatics 67, 17–33 (2014).
13. Balteiro, I.: A reassessment of traditional lexicographical tools in the light of new corpora: sports Anglicisms in Spanish. International Journal of English Studies 11(2), 23–52 (2011).
14. Zenner, E., Speelman, D., and Geeraerts, D.: Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. Cognitive Linguistics 23(4), 749–792 (2012).
15. Núñez, N. E. E.: Anglicisms in CREA: a quantitative analysis in Spanish newspapers. In: Language design: journal of theoretical and experimental linguistics 18, 215–242 (2016).
16. Alex, B.: 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), European Language Resources Association (ELRA) (2008).
17. Furiassi, C. and Hofland, K.: The retrieval of false anglicisms in newspaper texts. Corpus Linguistics 25 Years On, 347–363 (2007).

18. Serigos, J.: Using distributional semantics in loan word research: A concept-based approach to quantifying semantic specificity of anglicisms in Spanish. International Journal of Bilingualism 21(5), 521–540 (2007).
19. Moreno F. F., Moreno S. A.: Configuración lingüística de anglicismos procedentes de Twitter en el español estadounidense. Revista signos 51(98), 382-409 (2018).
20. Lee, D.: Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In: ICML 2013 Workshop: Challenges in Representation Learning. pp. 1-6. (2013).
21. Shi, W., Gong, Y., Ding, C., et al.: Transductive semi-supervised deep learning sing min-max features. In Proceedings of the European Conference on Computer Vision. pp. 299-315. (2018).