EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis

David Vilares, Miguel A. Alonso, Carlos Gómez-Rodríguez
Grupo LyS, Departamento de Computación, Universidade da Coruña
Campus de A Coruña s/n, 15071, A Coruña, Spain
{david.vilares, miguel.alonso, carlos.gomez}@udc.es

Abstract

Code-switching texts are those that contain terms in two or more different languages, and they appear increasingly often in social media. The aim of this paper is to provide a resource to the research community to evaluate the performance of sentiment classification techniques on this complex multilingual environment, proposing an English-Spanish corpus of tweets with code-switching (EN-ES-CS CORPUS). The tweets are labeled according to two well-known criteria used for this purpose: *SentiStrength* and a *trinary scale* (*positive, neutral* and *negative* categories). Preliminary work on the resource is already done, providing a set of baselines for the research community.

Keywords: Sentiment Analysis, Corpus Generation, Code-Switching

1. Introduction

Sentiment analysis (SA) is the field of research that deals with the automatic comprehension of the subjective information shared by users, especially on the Web (Pang and Lee, 2008; Cambria et al., 2013). One of its main challenges is *polarity classification*, focused on classifying a text, sentence or phrase as positive, negative or neutral (or even considering different levels of granularity like strongly positive and strongly negative (Nakov et al., 2016)). The interest of organizations and companies in this task has increased in recent years, due to the rise of social media. In particular, Twitter has become one of the most useful social networks for trending analysis, thanks to its abilities to capture popular trends and the easy interaction among its members.

SA techniques have been successfully applied on Twitter to monitor a wide variety of issues ranging from the perception of the public with respect to popular events (Thelwall et al., 2011) to real-time political analysis (Vilares et al., 2015d). Some of these trends are global (e.g. the Oscars, Superbowl or Rihanna) and so their trending topics are global too (e.g. '#oscars2016', '#superbowl2016', ...). However, the public perception of these trends often changes from one country to another and the task becomes even harder when tweets are written in different languages.

This has motivated the need of multilingual sentiment analysis. Usually, researchers evaluate multilingual approaches by translating (Balahur and Turchi, 2012) or merging monolingual corpora in different languages (Vilares et al., 2015c). But there exist cases where these synthetic corpora are not adequate to evaluate more difficult and unexplored multilingual variants, such as code-switching texts (i.e. texts that contain terms in two or more different languages). Colloquial creole languages such as *Spanglish* (a mix of Spanish and American English) or *Singlish* (English-based creole from Singapore) or even official languages such as the *Haitian creole* (which merges Portuguese, Spanish, Taíno, and West African languages), are some of the best-known situations of spoken code-switching.

The aim of this paper is to provide a resource to the research community that can be used to evaluate the performance of sentiment classification techniques on this complex multilingual environment, proposing an English-Spanish corpus of tweets with code-switching (EN-ES-CS CORPUS). To the best of our knowledge, this is the first code-switching collection annotated with sentiment labels.

The remainder of the paper is organized as follows. Section 2. describes the corpus. Section 3. shows preliminary research on the corpus, providing baselines for the natural language processing (NLP) community. Section 4. presents available NLP tools for multilingual and code-switching NLP. Finally, Section 5. draws our conclusions.

2. Corpus creation

To create the corpus, we take as starting point the collection presented in (Solorio et al., 2014), a workshop on language detection on code-switching tweets, where the goal was to apply language identification at the word level. The organisers proposed four code-switching language detection challenges: Spanish-English, Nepali-English, Mandarin-English and Modern Standard Arabic-Arabic dialects. They made the training corpora available to the research community, together with a small tuning collection, but no test set was released.

For building our resource, we just considered the Spanish-English training set (originally 11 400 tweets). As a first step, we removed all the non code-switching texts, i.e. those where all the words belonged to the same language, obtaining a filtered collection of 3 062 tweets. A number of different types of tweets can be found in the corpus:[1]

- Tweets that show (even opposite) sentiment in both languages, e.g. *'Tan bien que ivan las cosas... im so lost what did i do?!'*.

[1]The double underline represents the English text, the simple underline the Spanish phrases, and no underline illustrates language-independent symbols.

- Tweets where the sentiment is just in the English side of the tweet, e.g. *'I legitally screamed!!!! No fue una si no dos!!! '*.
- Tweets where the sentiment is just in the Spanish side of the tweet, e.g. *'This house da miedo'*.
- Tweets where the sentiment relies on language-independent symbols, e.g. *'Wow no lo puedo creer? -.-'*.

Those tweets were sent to three annotators fluent both in Spanish and English, who were asked to annotate them according to the SentiStrength criteria (Thelwall et al., 2010) and the Wiebe et al. (2005)'s annotation style. SentiStrength is a dual-score sentiment labeling strategy where each text is given two scores between 1 and 5: one indicating the positive strength (*ps*) of the tweet and the second one indicating its negative strength (*ns*). For example, *'I love you, but I hate you'* would have both a strong positive and negative sentiment. For inter-annotator agreement we relied on Krippendorff's alpha coefficient (Hayes and Krippendorff, 2007), obtaining an agreement from 0.629 to 0.664 for negative sentiment and 0.500 to 0.693 for positive sentiment.
Table 1 shows the frequency distribution of the SentiStrength scores and how annotators tend to often find slight levels of subjectivity, while highly subjective tweets tend to be less frequent.[2]

Positive	%tweets	Negative	%tweets
1	63,26	1	69,42
2	26,58	2	19,59
3	7,54	3	8,43
4	2,35	4	2,15
5	0,26	5	0,04

Table 1: Frequency distribution of the SentiStrength scores on the EN-ES-CS CORPUS

The results are coherent with other corpora annotated according to these criteria (Thelwall et al., 2010; **?**). The corpus was observed to be especially noisy, with many grammatical errors occurring in each tweet. Additionally, a predominant use of English was detected. We believe this is because the Solorio et al. (2014) corpus was collected by downloading tweets posted by people from Texas and California, where English is the primary language. Table 2 reflects these particularities.[3] In total, our collection contains 24 758 English terms, with 5 565 unique words, where 3 576 of them turned out to be out-of-vocabulary (OOV). Spanish is the minority language in the corpus, with 16 174 occurrences of terms and only 5 033 unique words, although with a larger percentage of OOV words. We also ran a language detection system, `langid.py`, resulting in 59.29% of tweets being predicted as English tweets.
Finally, there is also a nearly ubiquitous use of subjective clauses and abbreviations, especially *'lol'* and *'lmao'*, whose sentiment was considered a controversial issue by the annotators. It is interesting to point out that the presence of these clues was also used sometimes as a part of a negative message (i.e. *'He is so stupid, lmao'*), without any positive connotation. We believe this could have been one of the reasons why the inter-annotator agreement was lower for positive than for negative scores.

Language	Word occurrences	Unique words	OOV words
English	24 758	5 565	3 576
Spanish	16 174	5 033	3 714

Table 2: Word statistics by language on the EN-ES-CS CORPUS. Symbols like numbers or punctuation marks were considered language independent by (Solorio et al., 2014)

Table 3 shows some of the most common terms observed in our corpus that usually have sentiment associated, confirming the tendency of the users to employ subjective interjections coming from English. It is also important to note that the Spanish terms usually involve Mexican Spanish varieties, so specific resources from these might be needed to improve performance on the Spanish phrase sentiment classification.

English term	Occ.	Spanish term	Occ.
lol	474	bien	61
like	170	jajaja	29
lmao	122	mejor	28
haha	67	pinche	25
good	64	quiero	22
love	47	kiero	19
shit	47	jaja	18
fuck	42	guey	15
better	29	pedo	14

Table 3: Number of occurrences of some of the most common subjective terms for English and Spanish in the code-switching corpus

2.1. Additional labeling

A second labeling strategy is also provided for the code-switching corpus. After averaging the annotator scores, we applied a transformation to the *de facto* standard polarity classes (positive, neutral and negative) (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015). If $ps > ns$ then the tweet was considered *positive*. If $ps < ns$ then the tweet was considered *negative*. Otherwise, it was taken as *neutral*.[4] After the conversion, we obtained a collection where the *positive* class represents 31.45% of the corpus, the *negative* one represents 25.67% and with a 42.88% of

[2] Words such as *'good'* or *'bad'* tend to be more often used than *'spectacular'* or *'horrible'*, that are reserved for more special occasions.

[3] The words present in McDonald et al. (2013)'s English and Spanish treebanks were taken as our dictionaries. To know the language of each word of the corpus, we rely on Solorio et al. (2014)'s annotations.

[4] Neutral tweets can be either totally objective or mixing positive and negative sentiment with the same strength. However, the latter case turned out to be very uncommon.

neutral tweets. This frequency distribution is also close to that of other widely used Twitter corpora (Rosenthal et al., 2014). Both versions of the EN-ES-CS CORPUS can be obtained at `http://www.grupolys.org/software/CS-CORPORA/` or by asking any of the authors. We have tagged the corpus following different strategies in order to provide a richer resource, giving users the opportunity to select the tagging scheme that best suits their needs.
The format of the corpus labeled according to SentiStrength is:

`ps \t ns \t tweetid \t text`

and the format of the corpus labeled according to the trinary scale is:

`polarity \t tweetid \t text.`

where for each tweet, `ps` refers to its positive strength, `ns` to its negative strength, `tweetid` to its unique identifier, `text` to its contents, `polarity` to its polarity class and `\t` is used to represent a tab character.

3. Application to Sentiment Analysis

The EN-ES-CS corpus was employed to perform experiments using state-of-the-art supervised models and sets of features (Vilares et al., 2015a) on the trinary annotated corpus (Vilares et al., 2015c). As the EN-ES-CS CORPUS is used here as test set, the proposed approaches were trained as follows:

- *Monolingual approaches*: Two monolingual models, one for Spanish (*es-model*) and one for English (*en-model*), were trained using the TASS 2014 (Villena-Román et al., 2015) and SemEval 2014 (Rosenthal et al., 2014) corpora, respectively. The aim was to provide a baseline with the performance that a purely monolingual model can achieve on code-switching texts.
- *Majority language detection approach (mld-model)*: An automatic language detection system (Lui and Baldwin, 2012) is used to determine which one language is dominant in the tweet (assuming that, intuitively, the language that has a bigger presence in the tweet would contain the sentiment of the sentence), to then run the corresponding monolingual model.
- *Purely multilingual approach (en-es-model)*: A supervised model is trained on the union of two monolingual corpora.

The sets of features used in the experiments were different bags of words composed of: words (W), lemmas (L) and psychometric properties (P) coming from Pennebaker et al. (2001). Additionally, we include models using bigrams and also more accurate supervised models combining linguistic information.
The experimental results show how the multilingual approach, which is the only one able to consider features coming from the two languages, is the one that obtains the best performance under a number of different feature combinations. In particular, Table 4 shows the accuracy obtained on

Features	Models			
	en	**es**	**mld**	**en-es**
Words (W)	**55.65**	47.65	52.74	54.87
Lemmas (L)	55.68	48.66	53.00	**56.37**
Psychometric (P)	53.04	43.63	50.69	**53.69**
Bigrams of W	54.31	47.45	51.67	**54.34**
Bigrams of L	55.03	48.92	52.16	53.63
Bigrams of P	**49.48**	40.46	46.08	46.86
Combined (W,P,T)	**59.18**	48.27	56.53	58.52
Combined (L,P,T)	58.55	49.67	56.07	**59.11**
Combined (W,P)	58.72	49.90	56.40	**58.82**
Combined (L,P)	58.85	50.82	56.07	**59.34**

Table 4: Accuracy (%) on the code-switching set

Features	Models			
	en	**es**	**mld**	**en-es**
Words (W)	**54.20**	45.20	51.62	54.10
Lemmas (L)	54.30	46.20	51.89	**55.70**
Psychometric (P)	52.20	40.80	50.01	**53.30**
Bigrams of W	49.30	45.10	48.52	**51.90**
Bigrams of L	50.10	46.40	49.08	**51.40**
Bigrams of P	**47.70**	37.30	45.20	46.80
Combined (W,P,T)	58.30	47.10	56.07	**58.52**
Combined (L,P,T)	57.70	48.90	55.63	**58.60**
Combined (W,P)	58.00	48.40	55.90	**58.82**
Combined (L,P)	58.20	49.30	55.59	**58.90**

Table 5: Micro-averaged F1 (%) on the code-switching set

the code-switching collection by: (1) the English monolingual model (*en*) , (2) the Spanish monolingual model (*es*) and the multilingual model (*en-es*). In a similar line, Table 5 shows the performance of the same models under the micro-averaged F1 measure. Again, the multilingual approach outperforms the rest of the models under most of the proposed features sets. This reinforces the need of multilingual models to properly analyze this kind of texts and the utility of the presented corpus for future research in this area.

4. NLP tools for code-switching texts

Together with this paper we make available both part-of-speech and dependency parsing models that are able to process English-Spanish code-switching texts (Vilares et al., 2015b), so the research community can use them to explore richer linguistic approaches. They can be downloaded from `http://grupolys.org/software/TAGGERS` and `http://grupolys.org/software/PARSERS` or by asking any of the authors. Figure 1 shows how these bilingual models work better than the corresponding monolingual models. More recently, Ammar et al. (2016) have also shown the utility of using harmonized treebanks for universal parsing.

5. Conclusions

We present the first code-switching Twitter corpus for multilingual sentiment analysis, composed of tweets that merge English and Spanish terms. Some initial experiments have

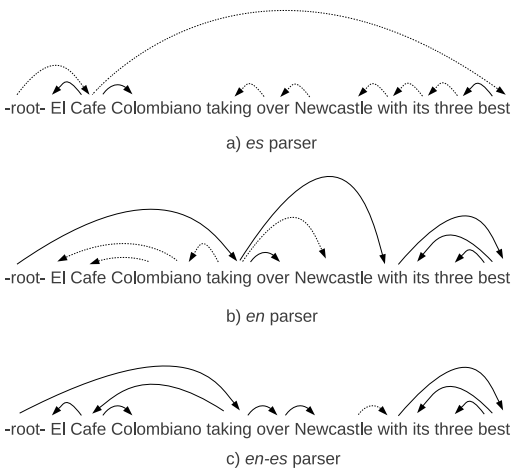

Figure 1: Example with the *en*, *es* and *en-es* dependency parsers. Dotted lines represent incorrectly-parsed dependencies

been already run, providing baselines for future research for the SA community. The results also show that neither monolingual nor multilingual approaches based on language detection are optimal to deal with code-switching texts, posing new challenges to sentiment analysis on this kind of texts.

6. Acknowledgments

This research is supported by the Ministerio de Economía y Competitividad (FFI2014-51978-C2) and Xunta de Galicia (R2014/034). David Vilares is funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180). Carlos Gómez-Rodríguez is funded by an Oportunius program grant.

7. Bibliographical References

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). One parser, many languages. *arXiv preprint arXiv:1602.01595*.

Balahur, A. and Turchi, M. (2012). Multilingual Sentiment Analysis using Machine Translation? In *WASSA 2012, 3rdWorkshop on Computational Approaches to Subjectivity and Sentiment Analysis, Proceedings of the Workshop*, pages 52–60, Jeju, Republic of Korea.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, (2):15–21.

Hayes, A. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97. Citeseer.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, June. ACL.

Nakov, P., Ritter, A., Rosenthal, S., Stoyanov, V., and Sebastiani, F. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. now Publishers Inc., Hanover, MA, USA.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, page 71.

Rosenthal, S., Nakov, P., Ritter, A., and Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of The 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014)*, pages 411–415.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, 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.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December.

Thelwall, M., Buckley, K., and Paltoglou, G. (2011). Sentiment in Twitter events. *J. Am. Soc. Inf. Sci. Technol.*, 62(2):406–418.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2015a). On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. *Journal of the Association for Information Science Science and Technology*, 66(9):1799–1816.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2015b). One model, two languages: training bilingual parsers with harmonized treebanks. *arXiv*, 1507.08449.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2015c). Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–8, Lisboa, Portugal, September. Association for Computational Linguistics.

Vilares, D., Thelwall, M., and Alonso, M. A. (2015d). The

megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *Journal of Information Science*, 41(6):799–813.

Villena-Román, J., Martínez-Cámara, E., García-Morera, J., and Jiménez-Zafra, S. M. (2015). TASS 2014-The Challenge of Aspect-based Sentiment Analysis. *Procesamiento del Lenguaje Natural*, 54:61–68.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.