A Multilingual Social Media Linguistic Corpus

Luis Rei[*,†], **Dunja Mladenić**[*,†], **Simon Krek**[*]
[*]Jožef Stefan Institute, [†]Jožef Stefan International Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: luis.rei@ijs.si, dunja.mladenic@ijs.si, simon.krek@ijs.si

Abstract

This paper focuses on multilingual social media and introduces the xLiMe Twitter Corpus that contains messages in German, Italian and Spanish manually annotated with Part-of-Speech, Named Entities, and Message-level sentiment polarity. In total, the corpus contains almost 20K annotated messages and 350K tokens. The corpus is distributed in language specific files in the tab-separated values format. It also includes scripts that enable to convert sequence tagging tasks to a format similar to the CONLL format. Tokenization and pre-tagging scripts are distributed together with the data.

Keywords: social media, Twitter, part-of-speech, named entities, named entity recognition, sentiment Analysis

1. Overview

High-quality newswire manually annotated linguistic corpora, with different types of annotations, are now available for different languages. Over the past few years, new social media based linguistic corpora have begun appearing but few are focused on classical problems such as Part-of-Speech tagging and Named Entity Recognition. Of these few, most are English corpora.

It has been documented that social media text poses additional challenges to automatic annotation methods with error rates up to ten times higher than on newswire for some state-of-the-art PoS taggers (Derczynski et al., 2013a). It has been shown that adapting methods specifically to social media text, with the aid of even a small manually annotated corpus, can help improve results significantly (Ritter et al., 2011; Derczynski et al., 2013a; Derczynski et al., 2013b).

While there exist social media sentiment corpora for twitter messages in the languages we annotated, the corpus we are presenting also includes message level sentiment labels. One motivation for this is the potential contribution of annotations, such as PoS tags, to sentiment classification tasks (Zhu et al., 2014).

The xLiMe Twitter Corpus provides linguistically annotated Twitter[1] social media messages, known as "tweets", in German, Italian, and Spanish. The corpus contains approximately 350K tokens with POS tags and Named Entity annotations. All messages, approximately 20K, are labeled with message level sentiment polarity. We further explain the composition of the corpus in § 3.

2. Related Work

An early effort in linguistically annotating noisy online text was the NPS Chat Corpus (Forsyth and Martell, 2007) which contains more than 10K online chat messages, written in English, manually annotated with POS tags.

The Ritter twitter corpus (Ritter et al., 2011) was the first to introduce a manually annotated Named Entity recognition corpus for twitter. It contains 800 English messages (16K tokens) which also contain Part-of-Speech and chunking tags.

The (Gimpel et al., 2011) corpus contains almost 2K twitter messages with POS tags while (Owoputi et al., 2013) annotated 547 twitter messages. Tweebank drawn from the latter boasts a total of 929 tweets (12,318 tokens) as well as providing clear guidelines which the previously mentioned twitter annotation efforts had not.

While there are many English social media sentiment corpora, the most well known is probably the Semeval corpus (Rosenthal et al., 2014) which contains over 21K Twitter messages, SMS, and LiveJournal sentences. All messages are annotated with one of three possible labels: Positive, Negative, or Objective/Neutral. For Spanish sentiment classification, the TASS corpus (Villena Román et al., 2013) contains 68K Twitter messages labeled semi-automatically with one of five labels: the three Semeval labels plus Strong Positive and Strong Negative. Smaller corpora with at least three labels exist for many other languages including German and Italian. We decided to add sentiment polarity to our multilingual corpus because it is a popular task, challenging for automated methods, and the cost (annotator time) of adding this additional annotation is mostly marginal when compared to the cost of PoS and NER annotations.

3. Description

The developed multilingual social media corpus includes document level and token-level annotations. There is one document level annotation, Sentiment polarity and two (2) token-level annotations, PoS and NER. The corpus details are shown in table 1, namely the distribution of annotated tweets and tokens per language. The Italian part of the corpus is the largest with 8601 annotated tweets, followed by Spanish with 7668 tweets, and German containing 3400 tweets.

3.1 Data Collection

The tweets were randomly sampled from the twitter public stream from late 2013 to early 2015. Tweets were selected based on their reported language. Some rules were automatically applied to discard spam and low information tweets ("garbage") tweets:

[1]Twitter: http://twitter.com

1. Tweets with less than 5 tokens were discarded;
2. Tweets with more than 3 mentions were discarded;
3. Tweets with more than 2 URLs were discarded;
4. Automatic language identification with langid.py (Lui and Baldwin, 2011) was used on the tweet text without twitter entities and if didn't match the reported language, the tweet was discarded.

Language	Tweets	Tokens	Annotators
German	3400	60873	2
Italian	8601	162269	3
Spanish	7668	140852	2

Table 1: Number of annotated tweets and tokens per language.

3.2 Preprocessing

URLs and Mentions were replaced with pre-specified tokens. Tokenization was performed using a variant of twokenize (O'Connor et al., 2010) that was additionally adapted to break apart apostrophes in Italian as in "l'amica" which becomes "l'", "amica".

3.3 Annotation Process

There were two annotators for Spanish, two for German, and three for Italian. A small number of tweets for each language were annotated by all the annotators working on the language in order to allow estimation of agreement measures as described in § 4.. POS tags were pre-tagged using Pattern (De Smedt and Daelemans, 2012) and some basic rules for twitter entities such as URLs and mentions.
We built an annotation tool optimized for document and token level annotation of very short documents, i.e. tweets. The annotation tool included the option to mark tweets as "invalid" since despite the automatic filtering performed in § 3.1 it was still possible that tweets with incorrectly identified language, spam, or incomprehensible text might be presented to the annotators. This feature can be seen in fig. 1.

3.4 Part-of-Speech

The part of speech tagset consists of the Universal Dependencies tagset (Petrov et al., 2012) plus twitter specific tags based on Tweebank (Owoputi et al., 2013). We present the full tagset and the number of occurrences, per language, of each tag in table 2.

3.41. Twitter Specific Tags

While most tags will be easily recognizable to most readers, we believe it is useful to provide here a description of the tags which are specific to social media and twitter. Further details about these tags can be found in our guidelines.

Continuation indicates retweet indicators such as "rt" and ":" in "rt @jack: twitter is cool" and ellipsis that mark a truncated tweet rather than purposeful ellipsis;

Figure 1: Screenshot of the annotation tool interface. The text of a tweet is at the top followed by the sentiment label dropdown menu. Below there is a column with the tokens and rows for each annotation (PoS and NER). Annotators manually fix the errors inherent in the automatic pre-tagging step previously described. Finally, a dropdown menu allows marking the annotation of the document as "To Do", "Finished", "Invalid", or "Skip". Note that in this example, the labels have not yet been manually corrected.

Tag	German	Italian	Spanish
Adjective	2514	7684	5741
Adposition	4333	14960	13467
Adverb	4173	8476	6116
Conjunction	1576	6737	6684
Determiner	2990	9811	10037
Interjection	225	1427	1109
Noun	11057	30759	23230
Number	1176	2550	1568
Other	1936	1503	3033
Particle	638	352	18
Pronoun	4530	7737	10333
Punctuation	8650	20529	14102
Verb	6506	21793	19460
Continuation	918	4227	3422
Emoticon	449	1076	951
Hashtag	1895	3035	1805
Mention	1984	6519	9070
URL	1923	4494	3019

Table 2: Tagset with occurrence counts in the corpus per language.

Emoticon this tag applies to unicode emoticons and traditional smileys, e.g. ":)";

Hashtag this tag applies to the "#" symbol of twitter hash-

tags, and to the following token if and only if it is not a proper part-of-speech;

Mention this indicates a twitter "@-mention" such as "@jack" in the example above;

URL indicates URLs e.g. "http://example.com" or "example.com";

A noteworthy guideline is the case of the Hashtag. Twitter hashtags are often just topic words outside of the sentence structure and not really part-of-speech. In this case, the Hashtag PoS tag applies to the word following the "#" symbol. Otherwise, if it is part of the sentence structure, the guideline specifies that it should be labeled as if the "#" symbol was not present.

3.5 Named Entities

Named entities are phrases that contain the names of persons, organizations, and locations. Identifying these in newswire text was the purpose of the CoNLL-2003 Shared Task (Tjong Kim Sang and De Meulder, 2003). We have adopted the definitions for each named entity class: Person, Location, Organization, and Miscellaneous. In table 3 we show each type of entity in our corpus and the number of tokens annotated with each per language.

Entity Type	German	Italian	Spanish
Location	742	2087	1441
Miscellaneous	995	5802	775
Organization	350	1150	836
Person	757	3701	2321

Table 3: Token counts per named entity type per language in the corpus.

3.6 Sentiment

Each tweet is labeled with its sentiment polarity: positive, neutral/objective, or negative. The choice of this three labels mirrors that of the Semeval Shared Task (Rosenthal et al., 2014). The vast majority of tweets in our corpus was annotated with the Neutral/Objective label as we show in table 4.

Language	Positive	Neutral	Negative	Total
German	334	2924	142	3400
Italian	554	7524	523	8601
Spanish	388	7083	197	7668

Table 4: Message level sentiment polarity annotation counts.

4. Agreement

In order to estimate inter-annotator agreement, for each language, the annotators were given tweets that they annotated in common. We show the number of tweets and tokens in table 5. These were then used to calculate Cohen's Kappa (technically, Fleiss' Kappa for Italian) and we show the results in table 6. The worst agreement between the human annotators occurred when labeling sentiment. Even for humans, it can be challenging to assign sentiment, without context, to a small message.

Language	Tweets	Tokens	Annotators
German	47	791	2
Italian	45	758	3
Spanish	45	721	2

Table 5: Number of tweets and tokens annotated by all annotators for a given language.

Task	German	Italian	Spanish
PoS	0.88 (AP)	0.87 (AP)	0.85 (AP)
NER	0.67 (SUB)	0.42 (MOD)	0.51 (MOD)
Sentiment	-0.07 (Poor)	0.02 (Slight)	0.37 (Fair)

Table 6: Inter Annotator Agreement (Cohen/Fleiss kappa) per task per language. In parenthesis, the human readable interpretation where: AP - Almost Perfect, MOD - Moderate, SUB - Substantial.

5. Format and Availability

The corpus is primarily distributed online[2] as a set of three tab-separated values (TSV) files - one per language. We also distribute the data in language and task specific formats such as a text file containing the German tweets with one word per line followed by a whitespace character and a NER label. These were automatically created using a script described in § 5.2.

5.1 Headers

Each of the TSV files has the same set of headers:

token the token, e.g. "levantan";

tok_id a unique identifier for the token in the current message, composed of the tweet id, followed by the dash character, followed by a token id, e.g. "417649074901250048-47407";

doc_id a unique identifier for the message (tweet id), e.g.: "417649074901250048";

doc_task_sentiment the sentiment label assigned by the annotator;

tok_task_pos the Part-of-Speech tag assigned by the annotator;

tok_task_ner the entity class label assigned by the annotator;

annotator the unique identifier for the annotator.

Note that the combination of the token identifier and the annotator identifier is unique i.e. the combination is present only once in the corpus.

[2]xLiMe Twitter Corpus: `https://github.com/lrei/xlime_twitter_corpus`

5.2 Scripts

In order to facilitate experiments using this corpus as well as to replicate its construction, several python scripts are distributed with the corpus data. We detail the most important scripts here. Namely the tokenizer, the pre-tagger, and the script that converts the sequence tagging tasks (PoS and NER) into a format similar to the CoNLL 2002/2003 format. In this format, there are empty lines which mark the end of a tweet and "word" lines start with the token followed by a space, followed by a tag.

xlime2conll.py the script used to convert the data into the column format similar to the CoNLL 2003 shared task;

extract_sentiment.py the script used to convert the data into a format that is easy to handle by text classification tools, specifically, a TSV file with the headers: id, text, sentiment;

twokenize.py the tokenizer used to split the tokens in the corpus;

pretag.py the script used to pre-tag the data;

agreement.py the script used to calculate the agreement measures.

6. Acknowledgments

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under xLiMe (FP7-ICT-611346) and Symphony (FP7-ICT- 611875). We would like to thank the annotators that were involved in producing the xLiMe Twitter Corpus. The annotators for German were M, Helbl and I. Škrjanec; for Italian, E. Dervišević, J. Jesenovec, and V. Zelj; and for Spanish, M. Kmet and E. Podobnik.

7. References

De Smedt, T. and Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Derczynski, L., Maynard, D., Aswani, N., and Bontcheva, K. (2013a). Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013b). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of Recent Advances in Natural Language Processing (RANLP).*, pages 198–206. Association for Computational Linguistics.

Forsyth, E. N. and Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 19–26. IEEE.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Lui, M. and Baldwin, T. (2011). Cross-domain feature selection for language identification. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*.

O'Connor, B., Krieger, M., and Ahn, D. (2010). Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM 2010)*, pages 384–385.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1524–1534.

Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Villena Román, J., Lana Serrano, S., Martínez Cámara, E., and González Cristóbal, J. C. (2013). Tass-workshop on sentiment analysis at sepln.

Zhu, X., Kiritchenko, S., and Mohammad, S. M. (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447.