A Multilingual Social Media Linguistic Corpus

Luis Rei[1,2] Dunja Mladenić[1,2] Simon Krek[1]

[1]Artificial Intelligence Laboratory
Jožef Stefan Institute

[2]Jožef Stefan International Postgraduate School

4th Conference on CMC and Social Media Corpora for the Humanities, September 2016

Table of Contents

Table of Contents

Tweets

German Italian Spanish

Manually Annotated with:

Part-of-Speech (PoS)

Named Entities (NE)

Sentiment

The xLiMe Twitter Corpus

Some Numbers

Language	Tweets	Tokens	Annotators
German	3400	60873	2
Italian	8601	162269	3
Spanish	7668	140852	2
Total	$\sim 20K$	$\sim 350K$	7

Motivation

Why We Created The Corpus

Evaluate the performance of NLP tools on tweets

- tweets are harder for automatic annotation methods (Derczynski, Maynard, *et al.* 2013)

Better Automated Methods (Ritter *et al.* 2011; Derczynski, Maynard, *et al.* 2013; Derczynski, Ritter, *et al.* 2013)

No Corpora available: German, Spanish, and Italian Twitter

- at least not with manually annotated NE and PoS

Sentiment classification with help from the other annotations

Cost vs Quality

A Note

Alternatives

- High Quality: N (e.g. N=7) expert annotators with full overlap
  - too expensive!
- Cheap: Crowdsourcing
  - time-consuming and/or low-quality

We opted for using language students as annotators. With the minimum possible overlap (while still possibly getting some measure of IAA).

Table of Contents

Related Work

Summary

- NPS Chat Corpus (Forsyth *et al.* 2007) [English]
  - $\sim 10K$ chat messages: PoS
- Ritter twitter corpus (Ritter *et al.* 2011) [English]
  - 800 tweets ($\sim 16K$ tokens): PoS, chunking tags, and NE
- Tweebank corpus (Owoputi *et al.* 2013) [English]
  - 929 tweets ($\sim 12K$ tokens): PoS
  - clear annotation guidelines and twitter-specific tagset
- Semeval 2014 Task 9 corpus (Rosenthal *et al.* 2014) [English]
  - $\sim 21K$ Twitter messages, SMS, and LiveJournal sentences
  - message level sentiment polarity: positive, objective/neural, negative

Table of Contents

Data Collection

Tweets were first selected based on their reported language.

The tweets were randomly sampled from the twitter public stream from late 2013 to early 2015.

Rules to **discard** spam and low information tweets tweets:

1. Less than 5 tokens;
2. More than 3 mentions;
3. More than 2 URLs;
4. Filtered using external automatic language identification tool [1]

[1] langid.py (Lui *et al.* 2011)

Preprocessing

1. URLs were replaced with pre-specified tokens;
2. @-Mentions were replaced with pre-specified tokens;
3. Tokenization
   - using a variant of twokenize (O'Connor *et al.* 2010)
   - adapted to break apart apostrophes in Italian
     - e.g. "l'amica" which becomes "l'", "amica"

Table of Contents

Annotation

Process

- $\sim$ 50 tweets were annotated by all the annotators working on the language
  - allows estimation of agreement measures
- Tokens pre-tagged with PoS labels [2]
  - Plus some basic rules for twitter entities such as URLs and mentions.
- Guidelines were sent to the annotators (now distributed with the corpus)
- A session was organised with all the annotators to explain the guidelines and the annotation software

[2] using Pattern (De Smedt *et al.* 2012)

Annotation Software

Description

We built an annotation tool optimised for document and token level annotation of tweets.

The annotation tool included the options to mark tweets as:

- invalid since despite the automatic filtering performed it was still possible that tweets with incorrectly identified language, spam, or incomprehensible text might be presented to the annotators
- skip in case they had doubts which added the tweet to a special list to which the annotator could come back later

in Dubai TURLURL Vergewaltigungsopfer mutmaßliches für Online-Petitition

Document Labels: sentiment: NEUTRAL

Token Labels:

Online-Petitition	pos: NOUN	ner: O
für	pos: ADPOSITION	ner: O
mutmaßliches	pos: ADJECTIVE	ner: I-ORG
Vergewaltigungsopfer	pos: NOUN	ner: I-PERSON
in	pos: ADPOSITION	ner: O
Dubai	pos: NOUN	ner: O
TURLURL	pos: URL	ner: O

ToDo
Finished
Invalid
Skip

← Previous | Document Index | Next →

Table of Contents

Part of Speech: Universal Tags

Tagset & Counts

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

Table: Tagset (Universal) with occurrence counts per language.

Part of Speech: Twitter-Specific Tags

Tagset & Counts

Tag	German	Italian	Spanish
Continuation	918	4227	3422
Emoticon	449	1076	951
Hashtag	1895	3035	1805
Mention	1984	6519	9070
URL	1923	4494	3019

Table: Tagset (Twitter) with occurrence counts per language.

Part of Speech: Twitter-Specific Tags

Description

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;
Emoticon	this tag applies to unicode emoticons and traditional smileys, e.g. ":)";
Hashtag	this tag applies to the "#" symbol of twitter hashtags, 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";

Named Entities

Tags & Counts

Named entities are phrases that contain the names of persons, organisations, and locations

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

Table: Token counts per named entity type per language

Sentiment

Labels & Counts

Each tweet is labeled with its sentiment polarity: positive, neutral/objective, or negative. The vast majority of tweets in our corpus was annotated with the Neutral/Objective

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

Table: Message level sentiment polarity annotation counts.

Table of Contents

Inter-Annotator Agreement

Overlap

In order to estimate inter-annotator agreement, for each language, the annotators were given tweets that they annotated in common.

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

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

Inter-Annotator Agreement

Results

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

The worst agreement between the human annotators occurred when labelling sentiment. Even for humans, it can be challenging to assign sentiment, without context, to a small message.

Table of Contents

Availability

Location & Content

The corpus is primarily distributed online:
`https://github.com/lrei/xlime_twitter_corpus`
under an Open Source License (MIT)

- 3 tab-separated values (TSV) files - 1 per language
- also task specific formats (e.g. CONLL)
- includes code:
  - tokenizer
  - pretag
  - agreement

Availability

File Format (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.

Thank You

Questions?

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).

Thank to 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.

References I

De Smedt, T. *et al.* Pattern for python. *The Journal of Machine Learning Research* (2012).

Derczynski, L., Maynard, D., *et al.* *Microblog-genre noise and impact on semantic annotation accuracy* in (2013).

Derczynski, L., Ritter, A., *et al.* *Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data.* in (2013).

Forsyth, E. N. *et al.* *Lexical and discourse analysis of online chat dialog* in (2007).

Lui, M. *et al.* *Cross-domain feature selection for language identification* in (2011).

O'Connor, B. *et al.* *TweetMotif: Exploratory Search and Topic Summarization for Twitter.* in (2010).

Owoputi, O. *et al.* *Improved part-of-speech tagging for online conversational text with word clusters* in (2013).

Ritter, A. *et al.* *Named Entity Recognition in Tweets: An Experimental Study* in (2011).

Rosenthal, S. *et al.* *SemEval-2014 Task 9: Sentiment Analysis in Twitter* in (2014).