A manually-annotated Italian corpus for fine-grained sentiment analysis

Marilena Di Bari, Serge Sharoff and Martin Thomas

DOI: 10.4000/books.aaccademia.1453
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: 11 November 2016
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899200008

Electronic reference
A manually-annotated Italian corpus for fine-grained sentiment analysis

Marilena Di Bari, Serge Sharoff, Martin Thomas
University of Leeds
School of Languages, Cultures and Societies
LS2 9JT, Leeds (UK)
{mlmdb,s.sharoff,m.thomas}@leeds.ac.uk

Abstract

English. This paper presents the results of the annotation carried out on the Italian section of the SentiML corpus, consisting of both originally-produced and translated texts of different types. The two main advantages are that: (i) the work relies on the linguistically-motivated assumption that, by encapsulating opinions in pairs (called appraisal groups), it is possible to annotate (and automatically extract) their sentiment in context; (ii) it is possible to compare Italian to its English and Russian counterparts, as well as to extend the annotation to other languages.

Italiano. Questo lavoro presenta i risultati dell’annotazione effettuata sulla sezione italiana del corpus “SentiML”, che consiste di testi sia originali che tradotti appartenenti a diversi tipi. I due vantaggi principali sono che: (i) il lavoro si fonda sull’assunzione motivata linguisticamente che, codificando le opinioni in coppia (chiamate appraisal groups), è possibile annotare (ed estrarre automaticamente) il loro sentimento tenendo in considerazione il contesto; (ii) è possibile confrontare l’italiano con le sue controparti inglese e russa, ed estendere l’annotazione ad altre lingue.

1 Introduction

Overall, the field of Sentiment Analysis (SA) aims at automatically classifying opinions as positive, negative or neutral (Liu, 2012). While at first the focus of SA was on the document level (coarse-grained) classification, with the years it has become more and more at the sentence level or below the sentence (fine-grained). This shift has been due to both linguistic and application reasons. Linguistic reasons arise because sentiment is often expressed over specific entities rather than an overall document. As for practical reasons, SA tasks are often aimed at discriminating between more specific aspects of these entities. For example, if an opinion is supposed to be on the plot of a movie, it is not unusual that the user also evaluates actors’ performance or director’s choices (Shastri et al., 2010). For SA applications these opinions need to be assessed separately. Also opinions are not expressed as simple and direct assertions, but by using a number of stylistic devices such as pronominal references, abbreviations, idioms and metaphors. Finally, the automatic identification of sarcasm, irony and humour is even more challenging (Carvalho et al., 2009).

For all these reasons, fine-grained sentiment analysis is looking at entities that are usually chains of words such as “noun+verb+adjective” (e.g., the house is beautiful) or “ad-verb+adjective+noun” (e.g., very nice car) (Yi et al., 2003; Popescu and Etzioni, 2005; Choi et al., 2006; Wilson, 2008; Liu and Seneff, 2009; Qu et al., 2010; Johansson and Moschitti, 2013).

In addition to the multitude of approaches to fine-grained SA, there is also shortage of multilingual comparable studies and available resources. To close this gap, we designed the SentiML annotation scheme (Di Bari et al., 2013) and applied it to texts in three languages, English, Italian and Russian. The proposed annotation scheme extends previous works (Argamon et al., 2007; Bloom and Argamon, 2009) and allows multi-level annotations of three categories: target (T) (expression the sentiment refers to), modifier (M) (expression conveying the sentiment) and appraisal group (AG) (couple of modifier and target). For example in:

“Gli uomini hanno il potere di [[sradicare]_M la [povertà]_T]_AG, ma anche di [[sradicare]_M le
(Men have the power to eradicate poverty, but also to eradicate traditions)
Table 1: Statistics on the annotated data. A different amount of appraisal groups has been annotated according to the text type, but on average the 24% of words are sentiment-loaded.

<table>
<thead>
<tr>
<th>Language</th>
<th>Text type</th>
<th>Appraisal groups</th>
<th>Targets</th>
<th>Modifiers</th>
<th>% of words included in appraisal groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITA</td>
<td>Political</td>
<td>486</td>
<td>411</td>
<td>437</td>
<td>25%</td>
</tr>
<tr>
<td>News</td>
<td>254</td>
<td>203</td>
<td>244</td>
<td>22%</td>
<td></td>
</tr>
<tr>
<td>TED</td>
<td>341</td>
<td>292</td>
<td>323</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>tot</td>
<td>1081</td>
<td>906</td>
<td>1004</td>
<td>24%</td>
<td></td>
</tr>
</tbody>
</table>

along with the percentages of words embedded in appraisal groups for each text type.

Figure 1 shows that ‘positive’ orientation is the predominant one for appraisal groups with 67%, followed by ‘negative’ with 32%. These data are consistent with the assumption that appraisal groups should not be ‘neutral’ nor ‘ambiguous’ because they carry appraisal and their orientation should be clear in context. At the same time, targets and modifiers can be ‘ambiguous’ because their orientation depends on the context and ‘neutral’ in case they are not the element carrying appraisal in the group.

Figure 2 shows the statistics on the other attributes: ‘appreciation’ is the most common attitude, which is consistent with the fact that this value is associate to ‘thing’ in the AF (see Section 1), which is the most common target type; polarity, which indicates that a negation has been encountered, has been ‘marked’ 4% times; force, an important feature for a more accurate prediction of the sentiment, is ‘reverse’ 4% of times.

We have also compared the contextual orientation manually annotated by us with the prior orientation included in the translation of the ‘positive’ and ‘negative’ values in the NRC Word-Emotion Association Lexicon (Mohammad, 2011), whose English annotations were manually done through Amazon’s Mechanical Turk, and the Roget Thesaurus and it has entries for about 14200 word types. We calculated that, in the case of Italian, only 29.39% of the words belonging to the appraisal groups were present in the sentiment dictionary, with higher percentage for political speeches (33.54%), followed by news (27.66%) and TED talks (26.98%). As previously found in the case of English, most of these are nouns reasonably not carrying sentiment on their own, but still part of an appraisal group (e.g., brevetti (patents), computer, confini (borders), nostro (our)). There are also cases, adjectives in particular, that should probably be included in a dictionary with prior orientation (e.g., necessario (necessary), negativo (negative), ovetato (overburdened), ideale (ideal)).

In line with our previous experiments in English (Di Bari et al., 2013), we used the following categories for the comparison:

**Agreeing words**: words whose dictionary orientation agrees with that of the appraisal group they belong to. They cover 69.63% of the total times words were found in the dictionary. This means that we can rely to a certain extent to the dictionary orientation, but not if we aim at more accuracy. The list includes reasonable out-of-context positive words (e.g., alleati (allies), comprensione (comprehension), dotati (gifted), felicità (happiness)), as well as out-of-context negative words (e.g., debolezza (weakness), malattia (sickness), stagnante (stagnant), violenza (violence)).

**Disagreeing words**: words whose dictionary orientation does not agree with that of the appraisal group they belong to. They cover 28.18% of the total times words were found in the dictionary, a percentage that demonstrates how crucial the context is. For example reversals such as abolire (abolish) and diminuire (diminish), and sfida (challenge), sopportare (to bear), tendenza (trend). However, it was interesting to notice that also words normally considered positive (e.g. prosperare (to prosper) and risorse (resources)) or negative (e.g. and tensione (tension) and rischio (risk)) became included in groups with opposite orientation.

**Ambiguous words**: words which already have both positive and negative values in the dictionary. They are resta (stays), rivoluzione (revolution), sciogliere (to unleash), umile (humble), and they cover 1.07%.
4 Conclusions

In this paper we have described a manually-annotated corpus of Italian for fine-grained sentiment analysis. The manual annotation has been done in order to include important linguistic features. Apart from extracting statistics related to the annotations, we have also compared the manual annotations to a sentiment dictionary and demonstrated that (i) the dictionary includes only 29.29% of the annotated words, and (ii) the prior orientation given in the dictionary is different from the correct one given by the context in 28.18% of the cases.

The original and annotated texts in Italian (along with English and Russian) and the Document Type Definition (DTD) of SentiML to be used with MAE are publicly available\textsuperscript{1}.

In the meanwhile, the authors are already working on an automatic system to identify and classify appraisal groups multilingually.

\textsuperscript{1}http://corpus.leeds.ac.uk/marilena/SentiML

Acknowledgments

The first author would like to thank Michele Filannino (The University of Manchester) for his insights throughout the research.

References


Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 100–107.


Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting
irony in user-generated contents: oh...!! it’s “so easy” ;). In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, TSA ’09, pages 53–56, New York, NY, USA. ACM.


Lokendra Shastri, Anju G. Parvathy, Abraham Kumar, John Wesley, and Rajesh Balakrishnan. 2010. Sentiment extraction: Integrating statistical parsing, semantic analysis, and common sense reasoning. In IAAI.


