Prior Polarity Lexical Resources for the Italian Language

Simone Faro [1], Valeria Borzì [1], Arianna Pavone [2] and Sabrina Sansone [2]

[1] Dipartimento di Matematica e Informatica, Università di Catania,
Viale A.Doria n.6, 95125, Catania, Italy
{faro,borzì}@dmi.unict.it

[2] Dipartimento di Scienze Umanistiche, Università di Catania,
Piazza Dante n.2, 95100, Catania, Italy
{pavone,sansone}@unict.it

Abstract. In this paper we present SABRINA (Sentiment Analysis: a Broad Resource for Italian Natural language Applications) a manually annotated prior polarity lexical resource for Italian natural language applications in the field of opinion mining and sentiment induction. The resource consists in two different sets, an Italian dictionary of more than 277.000 words tagged with their prior polarity value, and a set of polarity modifiers, containing more than 200 words, which can be used in combination with non neutral terms of the dictionary in order to induce the sentiment of Italian compound terms. To the best of our knowledge this is the first prior polarity manually annotated resource which has been developed for the Italian natural language.

1 Introduction

The preparation of manuscripts which are to be reproduced by photo-offset requires special care. Papers submitted in a technically unsuitable form will be returned for retyping or cancelled if the proceedings cannot otherwise be finished on time. Sentiment classification, described in Bing and Lei (2012), Liu and Zhang (2012) and Medhat et al. (2014), concerns the use of automatic approaches for predicting the orientation of subjective content on text documents, with applications on many areas including information retrieval, customer intelligence and recommender and advertising systems.

Such discipline, where sentiment, opinion or emotion, are identified and classified in human written text is well known as sentiment analysis.

With the rapid increase of available subjective text on the internet in the form of blog posts, comments in discussion forums and product reviews, mining the user's

opinion can assist in a lot of potential applications in areas such as recommender systems, search engines and market research.

Although some attempts have been made to extend solutions to other languages, till date all research efforts found in sentiment analysis literature deal mostly with English texts. However, in order to identify sentiment from a text, a lexical analysis of the source language plays a crucial role.

An approach for detecting sentiment in texts concerns the use of lexical resources such as a dictionaries of opinionated terms. For example the terms love , good and favorite directly indicate a positive sentiment or an opinion, while words like hate , bad and scandal can be associated with a negative sentiment.

Among the others, SentiWordNet, by Esuli and Sebastiani (2006), is one of the most used resource, containing opinion information on terms extracted from the WordNet database by Miller (1995) and made publicly available for research purposes. It is built via a semi supervised method and is considered a valuable resource for performing opinion mining tasks, providing a readily available database of term sentiment information for the English language.

Other previous works, as Pang and Lee (2002) and Esuli and Sebastiani (2006), have been already proposed for making dictionaries for those sentiment words using automatic approaches, however automatic identification of polarity orientation of such words is also a difficult research issue, known as polarity identification . In this context, it has been shown that the use of sentiment lexicons only provide a good baseline i.e. without using any natural language techniques only dictionary based approach produce a good performance, as noticed in Das and Bandyopadhyay (2010b).

An alternative to automatic tagged resources are manually annotated lexicons which turns out to be undoubtedly more trustable although they took long time to be constructed and may be subject it annotator bias.

In this paper we present SABRINA (Sentiment Analysis: a Broad Resource for Italian Natural language Applications) a manually annotated prior polarity lexical resource for Italian natural language applications in the field of opinion mining and sentiment induction. The resource consists in two different sets, an Italian dictionary of more than 277.000 words tagged with their prior polarity value, and a set of polarity modifiers, containing more than 200 words, which can be used in combination with non neutral terms of the dictionary in order to induce the sentiment of Italian compound terms. To the best of our knowledge this is the first prior polarity manually annotated resource which has been developed for the Italian natural language.

The paper is organized as follows. In Section 2 we introduce the concept of prior and posterior polarity and present some known lexicons which label terms with their sentiment polarity. Then in Section 3 we present the new tagged resources which has been created for the Italian language and discuss its properties. In Section 4 we briefly introduce also a web based fronted for accessing the resources. We draw our conclusions in Section 5.

2 Prior and Posterior Polarity

We would like to stress that the template should not be manipulated and that the guidelines regarding font sizes and format should be adhered to. This is to ensure that the end product is as homogeneous as possible.

A typical computational approach to sentiment analysis starts with prior polarity lexicons where entries are tagged with their prior out of context polarity as human beings perceive using cognitive knowledge.

The prior polarity of a term is the sentiment (positive or negative) that such word evokes by itself. More specifically we could define the prior polarity of a term as the polarity for its non-disambiguated meaning, out of any context.

For example the adjective cold evokes (in most cases) a fairly negative sentiment, since it is used in sentences as a cold man , a cold winter or I feel cold . However, depending on the context, we can find such term in sentences with a positive acceptation, as in I love cold beer.

In contrast with the prior polarity of a word, the polarities associated to each word sense is called in literature posterior polarity.

In most cases prior polarity lexicons are lists of positive and negative words, often deployed as baselines or as features for other methods for sentiment analysis research, as in Liu and Zhang (2012). In these lexicon, words are associated with their prior polarity. For example it is presumable that the term wonderful is associated with positive connotation while the term horrible is associated with negative one. These approaches have the advantage of not needing deep semantic analysis or word sense disambiguation to assign an affective score to a word and are domain independent. In other word they are less precise but more portable.

2.1 Polarity Lexicons

Opinion lexicons are resources that associate sentiment orientation and words. Their use in opinion mining research stems from the hypothesis that individual words can be considered as a unit of opinion information, and therefore may provide clues to document sentiment and subjectivity. These techniques could be broadly categorized in two genres: manual annotation and automatic extraction of word polarity.

Manual annotation. Manual annotated lexicons are undoubtedly trustable but it took long time and, for these reasons, tend to be constrained to a small number of terms. By its nature, building manual lists is a time consuming effort, and may be subject to annotator bias. Although such limitations manually created opinion lexicons were applied to sentiment classification as seen in Pang et al. (2002), where a prediction of document polarity is given by counting positive and negative terms.

Automatic detection. To overcome the above issues lexical induction approaches have been proposed in the literature with a view to extend the size of opinion lexicons from a core set of seed terms, either by exploring term relationships, or by evaluating similarities in document corpora. Early work in this area, by Hatzivassiloglou and McKeown (1997), extends a list of positive and negative adjectives by evaluating conjunctive statements in a document corpus. However in most cases automatic processes still demands manual validations and, moreover, may fail to cover the multiple domains as automatic processes trust on specific corpus.

SentiWordNet, by Esuli and Sebastiani (2006), is one of the most popular lexical resources in Sentiment Analysis. It has been widely adopted since it provides a broad-coverage lexicon, built in a semi-automatic manner, for English providing posterior polarities scores for each term of the language. It is the result of the automatic annotation of all the synsets of WordNet according to the notions of positivity, negativity, and neutrality. Different senses of the same term may thus have different opinion-related properties.

However in most opinion mining applications it is necessary to derive prior polarities starting from posterior polarities scores have been proposed in the literature. However, their performance varies significantly depending on the adopted variant. For instance SentiWords is an inducted prior polarity lexicon with the higher coverage for the English language. It contains roughly 155.000 words associated with a sentiment score included between -1 (strongly negative) and +1 (strongly positive), learned from SentiWordNet. Words in this resource are also aligned with WordNet lists. For the sake of completeness we notice also that other prior polarity sentiment lexicons are available for the English language, such as Subjectivity Word List, in Wilson et al. (2005), Word-Net Affect list, in Strapparava and Valitutti (2004), and the Taboada's adjective list, in Voll and Taboada (2007).

Although most of the efforts in literature have been devoted to the construction on lexicons resource for the English language, in recent years some research endeavors could be found in literature for solving the opinion mining problem in several languages and domains as in Das and Bandyopadhyay (2010b). Until date most of the approaches to sentiment analysis in languages different from English consists in applying a word-translation from the target language to English before polarity extraction, which is applied by using one of the above described lexicons. Such solutions, however, presents several problems including translation precision and disambiguation of words.

Recently some efforts have also been made to produce polarity lexicons for languages different from English. For instance Das and Bandyopadhyay (2010a) proposed multiple computational techniques like, WordNet based, dictionary based, cor-

pus based or generative approaches for generating SentiWordNet for Indian languages.

For the sake of completeness we mention also an interactive gaming approach used for obtaining polarity values of english words, presented by Das and Bandyopadhyay (2010b) who proposed a tool, named Dr. Sentiment, to create and validate SentiWordNet in 56 languages by involving Internet population.

3 New Broad Lexical Resources for the Italian Language

In this section we present SABRINA[1] (Sentiment Analysis: a Broad Resource for Italian Natural language Applications) a manually annotated prior polarity lexical resource for Italian natural language applications in the field of opinion mining and sentiment induction. The resource consists in two different sets, an Italian dictionary of more than 277.000 words tagged with their prior polarity value, and a set of polarity modifiers, containing more than 200 words, which can be used in combination with non neutral terms of the dictionary in order to induce the sentiment of Italian compound terms.

In recent years sentiment analysis in Italian texts has attracted attention due to Evalita, an initiative devoted to the evaluation of Natural Language Processing and Speech tools for Italian. In the recent Evalita 2014 edition the Sentipolc (SENTIment POLarity Classification) task[2] was proposed by Basile *et al*. (2014). It focused on Italian texts from Twitter by launching a battery of related tasks with an increasing level of complexity.

A first automatic annotated lexicon for the Italian language has been developed by **Basile and Nissim (2013)**, who exploited three existing resources, namely MultiWordNet by Ciravegna et al. (1994), SentiWordNet by Esuli and Sebastiani (2006), and WordNet, by Miller (1995), to obtain an annotated lexicon of senses for Italian.

It was named Sentix and basically port the SentiWordNet annotation to the Italian portion of MultiWordNet in a completely automatic fashion. Sentix was then used by Castellucci *et al*. (2014) who described the UNITOR system that participated to the Sentipolc task within the context of Evalita 2014.

The system has been developed as a workflow of Support Vector Machine classifiers. Specific features and kernel functions have been used to tackle the different subtasks, i.e. Subjectivity Classification, Polarity Classification and the pilot task Irony Detection. To the best of our knowledge, besides Sentix, SABRINA is the first prior polarity manually annotated resource which has been developed for the Italian natural language.

---

[1] A tool for evaluating SABRINA is available at the anonymous url http://www.dmi.unict.it/~faro/sabrina.

[2] http://www.di.unito.it/~tutreeb/sentipolc-evalita14/

Table 1. The distribution of polarity values assigned to Italian words.

polarity	value	# of words	% of words
strongly negative	-1.0	22.651	8.17%
negative	-0.5	49.074	17.70%
neutral	+0.0	162.170	58.47%
positive	+0.5	36.688	13.23%
strongly positive	+1.0	6.739	2.43%

3.1 Italian Polarity Lexicon

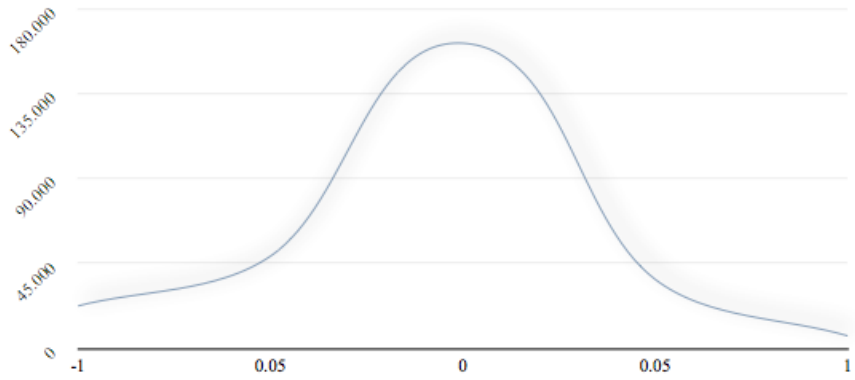
Most sentiment lexicons in literature contain lists of tagged lemmas, i.e. the canonical form (or dictionary form) of a word. For instance the lastest version of MultiWordNet (1.39) contains around 58,000 Italian word senses and 41,500 lemmas organized into 32,700 synsets aligned whenever possible with Princeton WordNet English synsets. In using such kind of resources in sentiment analysis it is necessary to operate a previous step of sense disambiguation in order to identify the correspondent lemma of a word.

Our lexicon contains 277,387 words of the Italian language, including their inflection, used in order to express different grammatical categories such as tense, mood, person, gender, etc. For instance the dictionary contains the verb correre (to run) and its conjugations correvo , correrà , corressi , etc.

Such set of words have been manually tagged with their prior polarities. The annotation process started from the word set in the Ispell Italian dictionary[3] used for spellchecking purpose. Each word of the lexicon has been associated with a polarity in the range between -1 and +1 , where -1 indicates a strongly negative polarity while +1 indicates a very positive polarity. Mildly negative or positive opinion polarity have been tagged, respectively, with values -0.5 and 0.5 . In addition terms with a neutral polarity have been tagged with a value equal to 0 .

---

[3] Ispell is a program that helps you to correct spelling and typographical errors in a file. When presented with a word that is not in the dictionary, ispell attempts to find near misses that might include the word you meant.



Two human annotators have been involved in the tagging process. The whole annotation process took more than three months.

Figure 1. The polarity distribution of the 277.387 different words of the Ispell Italian dictionary. Words are tagged with five different polarity values between -1.0 and +1.0.

Figure 1 shows the polarity distribution of all words of the Italian dictionary. We observed 162,000 words which have been tagged with a neutral sentiment polarity, more than 70,000 with a negative polarity and more than 43,000 words tagged with a positive polarity.

Specifically words evoking a negative sentiment are divided in two sets, 22,651 with a strongly negative polarity and 49,074 words with a fairly negative polarity. Similarly, in the case of words evoking a positive sentiment, we observed 6,739 words with a strongly positive polarity and 36,688 words with a fairly positive polarity. Table 1 shows in details the number of words detected for each polarity value together with the percentage of words detected in each group. Notice that more than 40% of words have been assigned to a polarity values, while 58% of words have been assigned with a neutral polarity.

2.3 Polarity Modifiers

An adjective is a word or set of words that modifies a noun or a pronoun. In most cases adjectives come before the word they modify. Some adjective can modify the polarity of a noun with a non neutral prior polarity. For example the adjective raro (rare) can be used in composition with the adjective bellezza (beauty) to emphasize its positive meaning (a women with a rare beauty). Similarly the adjective esiguo (scarse) can be used in combination with the noun valore (virtue) changing its positive polarity in a negative sentiment (a man with scarse virtue).

An adverb is a word or set of words that modifies verbs, adjectives, or other adverbs. Generally an adverb answers how, when, where, or to what extent an action is performed or an adjective is applicable. In this context some adverbs are able to modify the sentiment evoked by a verb or by an adjective with non neutral polarity. For instance the adverb appena (barely) can be associated with an adjective in order to reduce its positive (or negative) polarity, e.g. barely succeed or barely enthusiast .

Similarly the adverb davvero (truly) can be associated with an adjective like sorprendente (amazing) in order to emphasize its positive meaning.

In our work we collected a set of more than 200 polarity modifier which have been manually tagged with a proportionality factor ranging between -2.0 and +2.0 . When a term with a non neutral polarity x is associated with a modifier with a proportionality factor y , we obtain a compound term whose polarity can be estimated as (x y) . Depending on the value of such factor we can distinguish four different kind of modifiers.

Emphasize. These modifiers have a proportionality factor greater than +1.0 and, when associated with a term having a non neutral polarity, evokes a sentiment which is stronger than the original one. thus they emphasize a positive (or negative) polarity value.

proprio bello (really beautiful) = +1.6 +1.0 = +1.6
alquanto sgradevole (rather unpleasant) = +1.5 -1.0 = -1.5
grande valore (great virtue) = +1.8 +0.5 = +0.9

Moderate. These modifiers have a proportionality factor greater than 0 and smaller than +1.0. When associated with a term having a non neutral polarity, they result in a compound term with a moderated sentiment which is weaker than the original one.

appena vinto (just gained) = +0.7 +0.5 = +0.35
mediamente brutto (ugly on average) = +0.5 -1.0 = -0.5
breve successo (brief success) = +0.6 +0.5 = +0.3

Reverse and moderate. This kind of modifiers have a proportionality factor greater than -1.0 and smaller than 0.0 . When they are associated with a term having a non neutral polarity, evoke a sentiment which is in opposition with the original sentiment, but has an absolute value of polarity which is smaller than the original polarity.

poco ragionevole (little reasonable) = -0.7 +0.5 = -0.35
esiguo dolore (scarse pain) = -0.7 -1.0 = +0.7
limitato guadagno (limited benefit) = -0.8 +1.0 = -0.8

Reverse and emphasize. These modifiers have a proportionality factor smaller or equal than -1.0 and, if associated with a term having a non neutral polarity, evokes a sentiment which is stronger than the original one but with an opposite polarity.

insufficiente prestigio (insufficient prestige) = -1.2 1.0 = -1.2
minime scomodità (minimal inconvenience) = -1.0 -0.5 = +0.5
scarso valore (lacking virtue) = -1.2 +0.5 = -0.6

4 A Web Based Frontend

We implemented a simple web based tool in order to access the lexical resource presented in this paper. In order to allow a blind review of the paper we uploaded the tool in a free hosting server. The tool is accessible at the url

http://www.dmi.unict.it/~faro/sabrina

The tool allows to evaluate single Italian terms or compound terms, where words with a non neutral polarity are associated with modifiers, as described above. Moreover each example which you can find above in the paper is tagged with an anchor which redirect the reader to the web page of the tool in order to evaluate the sentiment value of the example itself. If a whole sentence is tested by the tool, containing more than one term with non neutral prior polarity, then a straightforward approach is applied in order to compute an approximation of the polarity of the whole sentence. In particular the set of polarity values contained in the sentence is arranged from the lowest one to the highest one and the median of such a set is taken as the polarity value of the whole sentence. Specifically the median is the number separating the higher half of the set of polarity values from the lower half. If there is an even number of polarity values, then there is no single middle value. Int this cases the median is usually defined to be the mean of the two middle values.

5 Conclusions

In this paper we presented a new lexical resource for the Italian language containing more than 277.000 words which have been manually tagged with their prior polarity values, i.e. a value indicating the sentiment which such words evoke when are out of any context. We also provide an additional lexical resource containing a set of more than 200 polarity modifiers which can be used for inducing the sentiment polarity of Italian compound terms. Future works will be devoted to test the effectiveness of such resource in opinion mining task.

References

Basile V. and Nissim M. (2013) Sentiment analysis on Italian tweets. In Proceedings of the 4th Ws: Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 100–107.

Basile V., Bolioli A., Nissim M., Patti V., and Rosso P. (2014) Overview of the Evalita 2014 SENTIment POLarity Classification Task. In Proceedings of the 4th evaluation campaign of NLP and Speech tools for Italian (EVALITA), Pisa, Italy.

Castellucci G., Croce D., De Cao D., Basili R. (2014) A Multiple Kernel Approach for Twitter Sentiment Analysis in Italian. In Proceedings of the First Italian Conference on Computational Linguistics (CLIC-IT).

Ciravegna F., Magnini B., Pianta E., Strapparava C. (1994) A Project for the Construction of an Italian Lexical Knowledge Base in the Framework of WordNet IRST Technical Report #9406-15.

Das A. and Bandyopadhyay S. (2010) SentiWordNet for Indian Languages. Proceedings 23rd International Conference on Computational Linguistics, pages 56–63.

Das A., Bandyopadhyay, S. (2010) Towards the Global SentiWordNet. Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation.

Bing L. and Lei Z. (2012) A Survey of Opinion Mining and Sentiment Analysis. Book Chapter, Mining Text Data, Springer US, pages 415–463.

Esuli A. and Sebastiani F. (2006) SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings ofLREC.

Hatzivassiloglou, V., and McKeown, K. (1997). Predicting the Semantic Orientation of Adjectives. Preceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL'97). Madrid, Spain, pp. 174-181.

Liu B. and Zhang. L. (2012) A survey of opinion mining and sentiment analysis. Mining Text Data, pages 415–463.

Medhat W., Hassan A., Korashy H., (2014) Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, Volume 5, Issue 4,1093–1113.

Miller G. A. (1995) WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Pang B., Lee L., and Vaithyanathan, S. (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of EMNLP.

Strapparava C. and Valitutti A. (2004) WordNet-Affect: an affective extension of WordNet. In Proceedings of LREC 2004, pages 1083–1086, Lisbon.

Voll K. and Taboada M. (2007) Not All Words are Created Equal: Extracting Semantic Orientation as a Function of Adjective Relevance. In Proceedings of the 20th Australian Joint Conference on Artificial Intelligence. pages. 337-346.

Wilson T., Wiebe J. and Hoffmann P. (2005) Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of HLT/EMNLP 2005.