An Approach to Detect Polarity Variation Rules for Sentiment Analysis

Pierluca Sangiorgi[1,2], Agnese Augello [1] and Giovanni Pilato[1]
[1] *ICAR, Istituto di Calcolo e Reti ad Alte Prestazioni, CNR, Consiglio Nazionale delle Ricerche,*
Viale delle Scienze - Edificio 11, 90128, Palermo, Italy
[2] *INAF, Istituto di Astrofisica Spaziale e Fisica Cosmica - Palermo, via U. La Malfa 153, 90146, Palermo, Italy*

Keywords: Subjectivity Analysis, Sentiment Analysis, Opinion Mining, Machine Learning.

Abstract: Sentiment Analysis is a discipline that aims at identifying and extract the subjectivity expressed by authors of information sources. Sentiment Analysis can be applied at different level of granularity and each of them still has open issues. In this paper we propose a completely unsupervised approach aimed at inducing a set of words patterns that change the polarity of subjective terms. This is a very important task because, while sentiment lexicons are valid tools that can be used to identify the polarity at word level, working at different level of granularity they are no longer sufficient, because of the various aspects to consider like the context, the use of negations and so on that can change the polarity of subjective terms.

1 INTRODUCTION

In recent years, with the advent of blogs, forums, social communities, product rating platforms and so on, users have become more active in production of large amount of information. Starting from news as well as products reviews and any other information systems that allow on-line user interaction in terms of contents, users provide information through their contributions in the form of opinions, discussions, reviews and so on.

Companies and organizations are increasingly interested in this type of information because it can be used as knowledge resource for operations of market survey, political behavior and, in general, measurement of satisfaction.

Information produced by network users usually regards what is known as their ”private states”: their opinions, emotions, sentiments, evaluations and beliefs (Quirk et al., 1985) (Banea et al., 2011). The term subjectivity is usually used in literature as a linguistic expression of private state (Banea et al., 2011).

One of the latest and most challenging research task is the automatic detection of subjectivity presence in text, which is named subjectivity analysis. The task to also identifying, where possible, its polarity, by classifying it as neutral, positive or negative is defined in literature as sentiment analysis.

Sentiment analysis may be realized at several levels of granularity, like word, sentence, phrase or document level. Usually each level of analysis exploits the results obtained by the underlying layers.

What makes these tasks hard to achieve is the ambiguity of words, the context-sensitivity of subjective terms, the need of appropriate linguistic resources for different languages, the presence of negations, the presence of irony, and so on (Montoyo et al., 2012).

In this paper we propose an approach to automatically discover several sequential patterns (i.e. a sequence of tokens in the sentence) in a specific language that change the polarity of subjective words. The novelty of the approach is that it is completely unsupervised, requiring only the use of one single linguistic resource: a sentiment lexicon, which is specific for the language under consideration. This can be done in order to build a polarity variation detector to be used in sentiment analysis applications.

2 RELATED WORKS

In sentiment analysis, the comprehension of sentences polarities is a complex task, which involves the analysis of the composition of the words in the sentences, considering their prior polarities. Several sentiment lexicons (Wilson et al., 2005a), (Stone and Hunt, 1963), (Baccianella et al., 2010), (Strapparava and Valitutti, 2004), can be accessed in order to examine the polarity at a “word-level”; in some cases

Sangiorgi P., Augello A. and Pilato G..
An Approach to Detect Polarity Variation Rules for Sentiment Analysis.
DOI: 10.5220/0004961903440349
In *Proceedings of the 10th International Conference on Web Information Systems and Technologies* (WEBIST-2014), pages 344-349
ISBN: 978-989-758-024-6

these lexicons can be obtained trough machine learning approaches (which cluster terms according to their distributional similarity (Turney and Littman, 2003)), or by means of bootstrapping methodologies starting from few term-seeds (Banea et al., 2011)(Pitel and Grefenstette, 2008).

Polarity of words is highly dependent on the domain in which they are used, so that adjectives with positive polarity could have an opposite or a neutral polarity in another domain. Moreover, especially in the context of product reviews users often adopt abbreviations and idioms; therefore methods for the automatic creation of lexicons may be very useful in this context also to improve existing dictionaries.

At a ”sentence-level” it is necessary consider the composition of the word, with their prior polarities, into the phrase. Different compositional models have been proposed in (Yessenalina and Cardie, 2011), (Wu et al., 2011) and (Chardon et al., 2013). In (Hu and Liu, 2004) a set of adjective words (opinions words) is identified using a natural language processing method in order to decide the opinion orientation of a sentence: for each opinion word its semantic orientation is determined, and the opinion orientation is then predicted by analyzing the predominance of positive or negative words. In (Moilanen and Pulman, 2007) a composition model based on a syntactic tree representation has been proposed. Many other factors have to be considered, especially the presence of polarity influencers (Wilson et al., 2005b) (Polanyi and Zaenen, 2006). In (Wilson et al., 2005b) a study on the most important features for recognizing contextual polarity has been performed, and the performance of these features has been evaluated by using several different machine learning algorithms. In (Tan et al., 2012) it has been proposed an automatic approach to detect polarity pattern rules, based on the extraction of typed dependency bi-grams of sentences and the use of Class Sequential Rules (CSRs) to derive polarity class rules from sequential patterns within the bi-grams. In (Tromp and Pechenizkiy, 2013) is proposed an algorithm for polarity detection based on the use of different heuristic rules.

3 AN APPROACH TO DETECT POLARITY VARIATION RULES

The proposed approach is thought to work in an unsupervised manner. It can be run just one time in order to define the rules. A subsequent re-run of this approach is not required, excepted for an extension of the number of rules using other data sets as inputs. The process can be viewed as a black box that produces, given a set of documents as input, a set of rules in a specific structure that will be described below. The entire process can be decomposed in a sequence of steps as reported in figure 1.

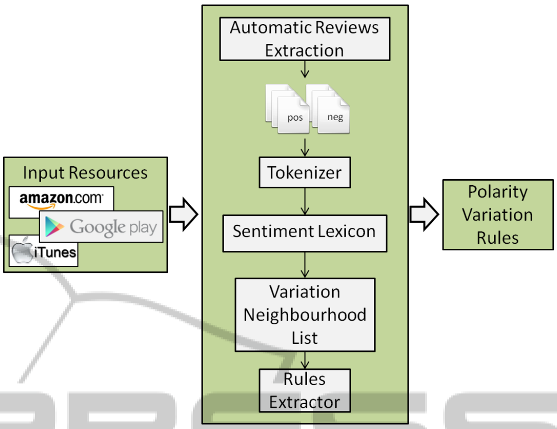

Figure 1: The architecture of the proposed approach.

The main idea is to extract several user reviews expressed in English language from an on-line market store which implements a star rating system. Each one of the reviews represents a short document, usually a sentence, that can be classified as having positive or negative polarity according to the number of stars that are compulsorily inserted by the users in order to publish a review.

Given these two classes of documents, after a simple pre-processing through a tokenizer used for splitting their content into words, we scan each word by exploiting a sentiment lexicon in English language. If in a document with a given polarity is detected a word with a different one, the three words to its left and those to its right are saved in a list.

Repeating this process for all the extracted reviews, we obtain a collection of words typically surrounding subjective words that, “for some reason”, are probably responsible for the variation of the polarity of the subjective words.

Without concerning why these rules change words polarity, we can study them in terms of frequency, support and confidence and select the most promising ones.

The collection of these sequential patterns will form a set of polarity variation rules. These rules could be integrated in a polarity variation detector in order to check, during other sentiment analysis applications, if a given sentence matches one of these new rules and therefore a different polarity for the term should be considered.

In a nutshell, the approach is based on four main steps:

1. extraction of a large amount of user reviews from on-line market store with star rating system;
2. sentiment analysis at word level using sentiment lexicons;
3. extraction of the left and right context of a subjective word with inverted polarity respect to that of the sentence;
4. identification of the sequential patterns rules that change the polarity;

It might seem that an approach like this could be replaced by a set of rules manually written by professional linguists, but this is not completely true. This because a linguist can define rules based on the grammar of the language that not always fits the "language" generally used by people on the social media. Furthermore, this approach can be potentially applied to any language of interest, since the polarity variation detection depends only on the tools used.

4 FORMALIZATION OF THE PROPOSED APPROACH

In our approach three main elements can be identified:

1. the input resources which consist of on-line user reviews that we want to analyze;
2. the processing chain;
3. the output which consist of linguistic patterns that cause the change of polarity of subjective terms.

We define reviews as input as an ordered sequence of variable length of words, punctuation and symbols z_i:

$$r_i = \{z_1, z_2, z_3, ..., z_n\}$$

and the desired output as two lists $P^+ = \{{p_i}^+\}$ and $P^- = \{{p_i}^-\}$, respectively for positive and negative sentences, of sequential patterns:

$$p_i = [n_1, n_2, n_3, n_4, n_5, n_6]$$

composed by an experimentally fixed number of terms, some of which empty, that represent the sequence of three terms that are present before (n_1, n_2, n_3) and after (n_4, n_5, n_6) a subjective term, which can cause its polarity variation.
The details of the entire processing chain will be described in the next subsections.

4.1 Acquisition and Data Preparation

The first step of the processing chain accomplishes the task of collecting reviews in order to create the knowledge resource, rich of subjective terms, from which to extract the polarity variation rules.

To this aim we consider the on-line user reviews as information sources, this because this type of text carries out opinions and sentiments expressed by the users, containing subjectivity terms with an high probability.

In particular, we are considering to retrieve the reviews from the Google Play market store, Amazon, Itunes or other on-line market shop.

This choice is motivated by a fundamental feature characterizing this kind of stores: they implement a star rating system, with scores usually from one to five, that must be inserted by the users in order to publish a review. This feature allows us to extract reviews as classified in terms of opinion (bad or good) expression. We make the supposition that reviews with higher value of stars express positive sentiments of users while reviews with lower value of stars express negative sentiments.

The result of this step is to extract several reviews with their associated stars values for different products and group them by positive or negative expressed opinion in two large class of documents $R^+ = \{{r_i}^+\}$ and $R^- = \{{r_i}^-\}$, considering only the stars values.
Once the classes of documents are built, they pass through a tokenizer chain to split the content of reviews in separated terms. The splitting must be done considering space character as well as punctuation, maintaining every remaining alphanumeric terms without filtering.

No other text pre-processing, like stopwords removal, must be done, this because we are interested to all terms that compose the reviews and could be determinant in the rules definition despite they not carry out particular information. For example think about terms like ”but”, ”not” and so on.

The result of this process is that every review is cleaned from spaces, punctuation and symbols. This lead to a review r_i composed only by terms t_i:

$$r_i = \{t_1, t_2, t_3, ..., t_n\}$$

4.2 Identification of the Variation Neighborhood Lists

Given the *i-th* review with positive polarity ${r_i}^+$, composed by a sequence of terms t_j:

$${r_i}^+ = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, ...\}$$

every *j-th* term t_j is checked through a sentiment lexicon, and whenever the term is detected as being subjective with negative polarity, the term is marked and its neighbor terms are stored in an array $N_i{}^+$ of predetermined size:

$N_i{}^+ = [n_{j-3}, n_{j-2}, n_{j-1}, n_{j+1}, n_{j+2}, n_{j+3}]$
where
$n = t$ (if t is a valid terms)
or
$n =$ "_" (if t is empty).

We define as "*Variation Neighborhood List*" a set of sequential patterns of fixed size composed by the terms in the left and in the right context of the subjective terms with discordant polarity, and in this case $N_i{}^+$ is an element of the Variation Neighborhood List $L^+ = \{N_i{}^+\}$ relative to the positive reviews.

This operation is executed for both positive and negative reviews. Generally speaking, we state that whenever a discordance between the prior polarity of a subjective term with the polarity of the review that the term belongs to is detected, a new element of the Variation Neighborhood List is built.

In order to clarify this statement, let us consider, for example, three negative reviews after the tokenizer process:

$r_1{}^-$ = {this, app, is, *good*, when, it, works, but, ninety, percent, of, the, time, it, will, not, even, open, on, my, phone}

$r_2{}^-$ = {my, previous, review, was, *good*, as, i, was, satisfied, with, the, app, but, now, feel, that, its, the, worst, app, ever}

$r_3{}^-$ = {not, *good*, needs, to, be, fixed}

All the three reviews contain the subjective term *good* that have positive prior polarity, which is however discordant with the overall sentiment expressed by the reviews.
In this case we build the three arrays:

$N_1{}^-$ = [this, app, is, when, it, works]

$N_2{}^-$ = [previous, review, was, as, i, was]

$N_3{}^-$ = [_, _, not, needs, to, be]

Note that the subjective term is discarded and not stored in the array, this because we are interested to its context which can affect its polarity as well as the context of other subjective terms.

The only restriction is to consider only an even number of terms for each context in order to have always patterns with the first half of terms referring to the left context of subjective terms and the second half to the right.

4.3 Rules Extractor

Once the two Neighborhood Lists L^+ and L^- respectively for positive and negative reviews are built, they are singularly processed in order to extract the two final set of polarity variation rules.

In order to reach this goal, the list is expanded adding all possible combinations of terms of the original sequential patterns extracted in the previews step. Considering the sequential pattern $N_2{}^-$ of the above example, this means create additional sequential patterns $N_{2,k}$ with k ∈ [1, 62], like:

$N_{2,32}{}^-$ = [previous, _, _, _, _, _]
$N_{2,16}{}^-$ = [_, review, _, _, _, _]
$N_{2,8}{}^-$ = [_, _, was, _, _, _]
$N_{2,39}{}^-$ = {previous, _, _, as, i, was}
and so on.

The *k-th* sequential pattern is built considering the binary encoding of k that leave the original terms in the positions with value equal to 1 and replace the terms with "_" in the positions with value equal to 0. k=0 and k=63 are not considered because they are respectively the empty pattern and the original pattern. Note that 63 is the maximum number of sequential patterns for each polarity variation detected, but they could be less if the original pattern have empty elements ("_").

After the expansion process, the next step is to compare all the sequential patterns of the expanded list, remove all the duplicates and saving for each pattern its value of frequency.

Since a pattern with few empty elements is strong and representative of polarity variation, but not probable to occur many times, instead of its frequency we consider a weight value w associated to each pattern; w is defined as:

$w = (f\text{-}1)*l$

where l is the number of non empty elements of the pattern ($l \in [1, 6]$) and f is the frequency of the pattern in the list.

In this way patterns that occur only once in the list have weight w equal to 0 and therefore they are discarded. Patterns with higher value of l, have an high confidence, but a very low support.

Considering only the frequency does not work: for example, let us consider that a pattern $N_{i,62}$ with a frequency value of 2, implies the existence of 6 patterns with the same frequency but l=1 ($N_{i,1}$, $N_{i,2}$, $N_{i,4}$, $N_{i,8}$, $N_{i,16}$, $N_{i,32}$), as well as the other "sub-patterns" with all value of l generated from the expansion procedure of the pattern with l=6.

Once the weight value w is calculated, for each patterns for the two Neighborhood Lists, using an appropriate value of threshold we remove all the patterns with a value lower than a given threshold .

All the patterns remaining from the cut-off phase represent the polarity variation rules we are looking for and stored as elements p_i in the two list of rules $P^+ = \{{p_i}^+\}$ and $P^- = \{{p_i}^-\}$.

They are expressed as vector of 6 terms, someone of which may be empty. These vectors, if used in sentiment analysis tasks, indicate, with a grade of probability, which terms must be present before and/or after a subjective term to cause its polarity variation.

5 CONCLUSIONS AND FUTURE WORK

We think that the work proposed in this paper could be a plausible approach to determine and solve the polarity variation problem in sentiment analysis applications. We are thus interested to develop this approach and we are working in this direction considering Google Play market store as resource for reviews and SentiWordNet, a lexical resource for opinion mining created by manual assignment of polarity to each synset of Wordnet, as sentiment checker. This will let us to define a set of valid rules to be integrated in a computer aided system for real time detection of polarity variation as support for other systems.

The choice to work in this direction, in fact, is not casual and arises from a real necessity occurred during other sentiment analysis works were we are involved. Just for example, during a sentiment analysis process done at word level on a large collection of Google Play market reviews, we have noticed that SentiWordNet found more positive than negative words inside reviews classified as negative with the star rating system. Find any discordant polarity terms inside a sentence it's a typical situation, but not in the numbers we found in our collection. This prompted us to investigate and define a valid approach to solve the problem.

ACKNOWLEDGEMENTS

This work has been partially supported by the POR 2007/2013 - Regione Siciliana - Misura 4.1.1.1. - IDS (Innovative Document Sharing) Research Project and by the PON01_01687 - SINTESYS (Security and INTElligence SYSstem) Research Project.

REFERENCES

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Banea, C., Mihalcea, R., and Wiebe, J. (2011). Multilingual Sentiment and Subjectivity Analysis. *ACL 2012*.

Chardon, B., Benamara, F., Mathieu, Y., Popescu, V., and Asher, N. (2013). Sentiment composition using a parabolic model. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 47–58, Potsdam, Germany. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Moilanen, K. and Pulman, S. (2007). Sentiment composition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, pages 378–382.

Montoyo, A., Martínez-Barco, P., and Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4):675 – 679.

Pitel, G. and Grefenstette, G. (2008). Semi automatic building method for a multidimensional affect dictionary for a new language. In *LREC*. European Language Resources Association.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In Shanahan, J., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10. Springer Netherlands.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman, London.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

Strapparava, C. and Valitutti, A. (2004). Wordnet affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Tan, L.-W., Na, J.-C., Theng, Y.-L., and Chang, K. (2012). Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. *Journal of Computer Science and Technology*, 27(3):650–666.

Tromp, E. and Pechenizkiy, M. (2013). Rbem: A rule based approach to polarity detection. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 8:1–8:9, New York, NY, USA. ACM.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005a). Opinionfinder: A system for subjectivity analysis. In *HLT/EMNLP*. The Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2011). Structural opinion mining for graph-based sentiment representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1332–1341, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yessenalina, A. and Cardie, C. (2011). Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 172–182, Stroudsburg, PA, USA. Association for Computational Linguistics.