A SEMANTIC APPROACH TO THE EXTRACTION OF FEATURE TERMS

Manuela Angioni and Franco Tuveri
CRS4, Center of Advanced Studies, Research and Development in Sardinia, Parco Scientifico e Tecnologico Ed. 1, 09010, Pula (CA), Italy

Keywords: Sentiment analysis, Opinion mining, NLP, Text categorization, Semantic disambiguation.

Abstract: Understanding the meaning of a text depends on the knowledge the reader has about the topic addressed in a document, starting from the most complex concept to the simplest one. The representation of the knowledge is generally performed by ontologies, semantic networks, or typified by statistical algorithms able to organize the contents according to rules based on frequency of terms or synsets. The Opinion Mining is a way to go beyond text categorization through the analysis of the opinions related to a specific topic: a product, a service, a tourist location, etc. In this paper we propose to apply our experience in the semantic analysis of textual resources to the Opinion Mining task, with the aim to propose a different approach to the extraction of feature terms, performing a contextualisation by means of semantic categorisation, a semantic net of concept and by a set of qualities associated to the sense expressed by adjectives and adverbs.

1 INTRODUCTION

Nowadays the digital portion of our life is constantly increasing and the way people interact is evolving from a restricted face-to-face friendly relation to a wider Web social one.

Currently millions of people use regularly social networks to communicate and to share information, interests and activities. They express opinions about any topic, talk about their lifestyles, needs or wishes, and the *user generated content* expressed in form of reviews of product on blogs, forum, discussion groups, is growing rapidly, becoming a valuable resource.

A such amount of social data can be used to analyze the present and to predict the near future needs or the probable changes. Reviews are used every day by common people or by companies who need to make decisions. They facilitate to book an hotel or a restaurant, to buy a book, or to taste the market tracing the customer satisfaction about a product. Mining the opinions and the comments is a way to extract knowledge by previous experiences and by the received feedback.

As showed in Figure 1, the Hype Cycle, from a Gartner analysis for the 2010 year, represents graphically the expectations about emerging technologies. The need for automated methods is growing and social analytics offers an answer (Crimson Hexagon, 2009), as one of the key themes emerging in the near future.

It is evident that even the opinion monitoring is essential for listening to and taking advantage of the conversations of possible customers in a data-driven decision making process.

Currently the main objective is to move forward the frontier of the Opinion Mining passing from the simple evaluation of the polarity of the expressed feeling, to a one where sentiments extracted are context related and the information related to the features is more detailed.

The main goal of our work is the development of an Opinion Mining system able to extract the features and the meaningful information related to them and contained in opinions, in a general and not clearly defined domain, from multiple review sources.

The term feature is here used with the same sense given by (Ding et al., 2008) in their approach to the Opinion Mining. It seems of considerable interest to explore the issue of the contextualization of features from a review in order to more easily combine the qualitative information expressed by the users to the highlighted features.

Another relevant aspect is the fact that people often use different terms to describe the same

Angioni M. and Tuveri F..
A SEMANTIC APPROACH TO THE EXTRACTION OF FEATURE TERMS.
DOI: 10.5220/0003602204020407
In *Proceedings of the 6th International Conference on Software and Database Technologies* (ICSOFT-2011), pages 402-407
ISBN: 978-989-8425-77-5

feature. The use of synonyms allows us to extract significant features from a text by a disambiguation analysis. Sometimes it happens that a feature could be related both to the specific domain and to the application and that some features that suggest the same concept are not expressed in WordNet (Miller, 1998) as synonyms.

As in (Benamara et al., 2007), we propose a linguistic approach to Opinion Mining, based on a combination of adverbs and adjectives, but our work introduces the use of the WordNet synsets related to each term. Our approach focuses on the analysis of the opinions through the processing of the textual resources, the information extraction and the evaluation of a semantic orientation. More in details, it performs a semantic disambiguation and a categorization phase and takes into account the meaning express in conversations, considering for instance the synonyms related to relevant terms. Moreover some properties related both to adjectives and to adverbs are associated to each synset, providing the result of the analysis grouped into thematic categories.

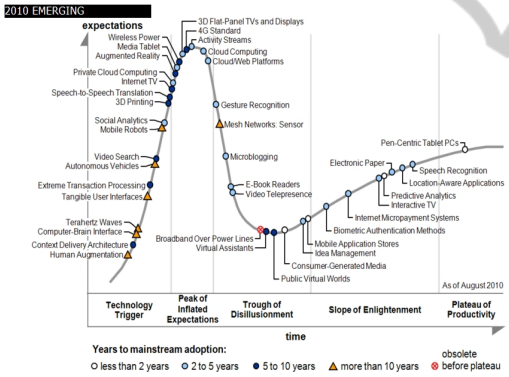

Figure 1: Hype Cycle for Emerging Technologies 2010.

The remainder of the paper is organized as follows: Section 2 refers to related works. Section 3 introduces our approach and examines the work performed on adjectives' and adverbs' structures and gives some details about the semantic classifier and the SemanticNet. Section 4 explains our approach to the feature extraction. Finally, Section 5 draws conclusions.

2 RELATED WORKS

Many Opinion Mining approaches are based on the use of lexicons of words able to express subjectivity, without considering the specific meaning the word assumes in the text by means of any form of semantic disambiguation.

Other approaches consider instead the word meanings as (Akkaya et al., 2009), that build and evaluate a supervised system to disambiguate members of a subjectivity lexicon, or (Rentoumi et al., 2009), that propose a methodology for assign a polarity to word senses applying a Word Sense Disambiguation (WSD) process.

As asserted by (Lee et al., 2008), "Opinion Mining can be roughly divided into three major tasks of development of linguistic resources, sentiment classification, and opinion extraction and summarization". The task of the development of linguistic resources could be performed by several methods: the conjunction method (Hatzivassiloglou and McKeown, 1997), the Pointwise Mutual Information (PMI) method (Turney et al., 2002), the WordNet exploring method, and the gloss classification method in the development of linguistic resources. In the sentiment classification phase some researcher team (Pang et al., 2002) applied standard machine learning classification techniques for classifying the sentiment related to the opinion of a document with a polarity classification, positive/negative, as a special case of text categorization with sentiment - rather than topic - based categories. Talking about the methods used in the opinion summarization step, in (Lee et al., 2008) is asserted that the introduction of the sense disambiguation in text analysis showed that systems that adopt syntactic analysis techniques on extracting opinion expressions tend to show higher precision and lower recall than those which do not adopt this kind of techniques. One relevant task in opinion summarization step regards the features extraction. Some works about features are based on the identification of nouns through the pos-tagging and provide an evaluation of the frequency of words in the review based on tf-idf criterions and its variation (Scaffidi, 2007). Others (Zhai et al., 2010) propose a constrained semi-supervised learning method to solve the problem to group feature expressions while (Popescu and Etzioni, 2005) worked on the explicit features in noun phrases.

Regarding the topics discussed above, the two following works seem interesting and have in part inspired our studies. The first is SentiWordNet (Esuli and Sebastiani, 2007) in which WordNet is expanded thanks to a semi-automatic acquisition of the polarity of WordNet terms, evaluating each synsets according to positive, negative and objective values. The other work (Zhai et al., 2010), as said previously, proposes a method based on the contextualization of reviews grouped on specific

domains and on a characterization based on features defined by users, trying in such a way to solve the problem to group feature expressions and to associate them to feature labels. They do not use WordNet that is not considered sufficient for several reasons, including the problem of semantic disambiguation, the lack of technical terms or specific meanings related to the context of use, or yet the differences of synonymy between different context.

3 OUR APPROACH

The aim of our research activity is to develop and build tools able to extract information and meaning from the analyzed texts. In this paper we illustrate some of these tools and how the Opinion Mining tasks may benefit by their use, in particular in order to propose an approach to extract and contextualize features related to products or services.

According to the steps about Opinion Mining, an overview about the activities we are working on is given in the following.

Syntactic and Semantic Analysis: pos tagging, adjectives and adverbs extraction and Word Sense Disambiguation activities.

Sentence Analysis: distinction between objective and subjective phrases and extraction of information by means of adjectives qualities and previous syntactic analysis; sentence categorization.

Document Analysis: context definition of reviews by means of the semantic classifier through syntactic and semantic analysis and the SemanticNet; document categorization.

Features Extraction: features extraction from sentences; distinction between implicit and explicit features by means of semantic analysis; analysis of noun phrases; analysis and validation of the features extracted; assignment of a weight to each feature.

The result of this analysis could be used in a summarization step in order to define the polarization of reviews related to the features extracted.

On the following we describe some tools and lexical resources we developed during our last research activities that we aim to apply in the steps of extraction of meanings and features from textual resources.

a. Adjectives and Adverbs Qualities

One of the goals of this work is the building of a lexical database of synsets enriched with a set of properties related to some of the adjectives and adverbs available in WordNet, with a positive, negative or neutral value associated. The addition of information given by the qualities associated to each synset can surely help to better identify the sentiment expressed in relation to the features and gives more details about the features and the result of the analysis certainly benefits from it.

Some linguistic resources are built considering three properties: subjectivity, orientation, and strength of term attitude. For example, ‘good’, ‘excellent’, and ‘best’ are positive terms while ‘bad’, ‘wrong’, and ‘worst’ are negative terms. ‘Vertical’, ‘yellow’, and ‘liquid’ are objective terms. ‘Best’ and ‘worst’ are more intense than ‘good’ and ‘bad’ (Lee et al., 2008).

Our analysis concentrates instead mainly on the qualitative adjectives able to specify for instance color, size, smell, making the meanings of sentences clearer or more exact. We have thus extended the properties of the semantic network of WordNet focusing on the characteristics of adjectives. We have classified about 2.300 pairs of adjectives/synsets and about 480 pairs of adverbs/synsets according to a set of attributes identified by their association with nouns and verbs and chosen on the basis of their frequency of use in the language. This work has been performed for three languages: Italian, English and Spanish using the data retrieved by FreeLing (Atserias et al., 2006); (Vossen, 1998) for English and Spanish and building *ex novo* a set of about 11.000 Italian terms, that in future will be made available freely online.

For each adjective, all the possible synsets available on WordNet has been considered and, for each of the meaning expressed by a synset, one of the previous categories has been associated, assigning a positive, negative or neutral value too.

The identified characteristics provide additional information about the content of the sentences, regarding for instance personal, moral, ethical or even aesthetical aspects. Some of these categories allow a polarization that can be used by Opinion Mining algorithms. In other cases it is immediately obvious that adjectives contain meanings intrinsically related to geographic, to time or to weather aspects. In our opinion, the use of such qualities associated both to adjectives and adverbs is useful to identify a first level of contextualization about objective and subjective phrases allowing to refer to things, people, places, weather conditions that can be contextualized on specific features.

Adverbs are useful too for the identification of the sentiment into the Opinion Mining process. We concentrate on some adverbs classified by their

meaning, they position or their strength, associating to each of them a specific synset as made for the adjectives. Based on their characteristics we have considered adverbs of manner, adverbs of place, adverbs of time, adverbs of quantity or degree, of affirmation, negation or doubt, adverbs as intensifiers or emphasizers and adverbs used in adversative and in consecutives sentences. Only the adverbs of manner may be positive, negative or neutral (objectives). The adverbs of degree give the idea about the intensity with which something happens or have an impact on sentiment intensity. The others give additional information to the analysis related to the location or the direction, the time.

The introduction of the synsets instead of considering only the words as keywords, extending in future work a similar evaluation to nouns and verbs, allows to have immediately the same qualities and values for the languages whose mapping between synsets is available.

b. The Semantic Classifier

Semantic text categorization techniques allow to classify texts contained in general text resources into one or more predefined categories, according to the meaning expressed in their content. The classifier we developed is composed by several modules able to perform syntactic parsing, word sense disambiguation and synsets extraction. A specific module implements a "density function" able to assign a weight to each synset extracted from the words in the sentences and in the document, and to each synset some coarse grain domains categories.

Through the semantic disambiguation task it is possible to reduce the number of synsets activated by the syntactic analysis. Calculating the synset density in a document we can take advantage of the semantic relations available from WordNet. Moreover, we can categorise sentences, assigning them some domain topics during a semantic disambiguation phase, identifying all the synsets referred to a textual content, and evaluating its most probable sense (Angioni et al., 2008a).

The classifier is capable to categorize text documents automatically, applying a classification algorithm based on the Dewey Decimal Classification, as proposed in WordNet Domains (Magnini et al., 2002); (Magnini et al., 2004), a lexical resource representing domain associations among the word-senses of WordNet. The results set of categories is further reduced by the application of a function that takes into account only categories characterized by a density value bigger than a fixed range value. To assess the performances of the classifier we selected five WordNet Domains categories: Animals, Plants, Medicine, Geography and Chemistry, and for each of them a set of 300 documents has been selected by hand. Table 1 and 2 summarize our results. More in details, for each category, the performances of the semantic classifier (MDC) are compared with a classifier that use a simple bag of synsets (BoS) of the terms of the document collection without any disambiguation. We resorted to different metrics aimed at evaluating precision π and recall ρ. In particular, we used micro- and macro-averaging, together with the F1 measure obtained by moving the acceptation threshold of each classifier under investigation over the range [0,1] (Sebastiani, 2002).

Table 1: Experimental results.

	π		**ρ**		**F1**	
	BoS	MDC	BoS	MDC	BoS	MDC
Animals	0,935	0,986	0,386	0,963	0,547	0,974
Plants	0.986	0,993	0,75	0,973	0,852	0,983
Medicine	0.939	1,00	0,67	0,973	0,782	0,986
Geography	0,676	0,970	0,953	0,986	0,791	0,978
Chemistry	0,924	0,983	0,903	0,98	0,913	0,981

Table 2: Micro and macro-averaging.

	π		**ρ**		**F1**	
	BoS	MDC	BoS	MDC	BoS	MDC
μAvg	0,857	0,986	0,732	0,975	0,790	0,981
MAvg	0,892	0,986	0,732	0,975	0,777	0,981

c. The SemanticNet

Mental associations of places, events, persons and things depend on the cultural backgrounds of the users' personal history. In fact, the way and the ability to associate a concept to another is different from person to person. In (Angioni et al. 2008b) a semantic net, based on the dictionary of WordNet and the contents included in Wikipedia, has been developed. Starting from the information contained in Wikipedia about a term of WordNet, the SemanticNet has been defined by adding new nodes, links and attributes, such as IS-A or PART-OF relations. Moreover, a classifier categorises the textual contents of Wikipedia and associates them to the WordNet synset having the most correct meaning by means of a similarity algorithm.

We therefore need to extract the specific meaning described in the Wikipedia page content, in order to build a conceptual map where the nodes are the "senses" (synset of WordNet or term+category)

and links are given both by the WordNet semantic relations and by the conceptual associations built by the Wikipedia authors.

Starting from each synset a new node of the conceptual map is uniquely identified by a semantic key (synsetID), by the term referred to the Wikipedia page and by the extracted categories. Through the information extracted from Wikipedia it is possible to build nodes, having a conceptual proximity with the beginning node, and to define, through these relations, a data structure linking all the nodes of the SemanticNet.

The semantic net, as it is defined, is not useful in this kind of application neither in the filtering of the features. We are planning the evaluation of a such data structure in order to improve the SemanticNet performances and to delimit the context of a feature, defined as a node in the net, by means of its relations (e.g. IS-A, PART-OF) and by its properties. So we are working to a way to increase its nodes and relations by means of some techniques of features extraction. The features extracted will be inserted in the graph of the semantic net extending it, whenever it will be possible, with new nodes and semantic relations inferred from the text analysis.

A recent version of the SemanticNet contains images as concepts related to its nodes. So, we are planning to investigate about the use of images as features referred to a specific object. In fact, in ever more web sites dedicated to reviews, frequently happens to find several images reporting objects and/or its features. An example is given by the pictures of a room or a bathroom in a hotel with the reviews associated.

4 FEATURES EXTRACTION

As said previously, the main objective of our researches is the extraction of meanings from texts applying linguistic approaches and semantic tools and resources, developed for general purposes, in Opinion Mining. The aspect we want to investigate is related to opinions about events or facts even in change where it is difficult to work on a set of reviews well defined or clearly related to a specific topic. An example could be related to politics where, in relation to an event of some kind, a politician reacts expressing his opinion or acting in some way. It could be very relevant investigate how such reaction is accepted by the electorate. Certainly in this case features could not be determined *a priori*, but should be extracted from text automatically. We have been inspired by the work done by (Yi, J. et al., 2003) and (Yi, J. et al., 2005) integrating their approach with some tools developed in our previous works, intervening in the phase of the extraction of the features, filtering the list of features through the semantic classification and through a contextualization phase by means of the semantic network.

In order to extract the features, we essentially make use of three tools: the semantic classifier, a categorisation tool based on quality categories associated to adjectives and adverbs, and the semantic net. The two tools of categorisation parse the phrases and allow to decide if and how a specific phrase is relevant in respect of a topic. They also allow to distinguish between objective and subjective sentences. More in details, the two classifier are complementary. The first classifier works only on nouns, while the other works on adjectives and adverbs, but until now we do not have any measure. Categories extracted by the classifier define a context, a topic, for the object in the review. For example, analysing reviews about tourism and especially reviews linked to hotels, we expect to examine sentences containing opinions about geographical locations, buildings, rooms, staff, food. This kind of information can be easily categorized by the classifier in categories that have a counterpart in adjective qualities. In fact, as said in Section 3.1, adjectives are defined with qualities related to places, geographical features, behavioural characteristics and thus related to people, measures, aesthetic qualities, and polarities for some of them.

All these aspects, although allow to highlight some additional information related to the subject of the review, are often not adequate in order to give a complete description of all features and attributes involved.

5 CONCLUSIONS

Several Opinion Mining methods and techniques have been developed in order to analyze contents and reviews. In this paper we proposed a different approach to the extraction of feature terms, performing a contextualisation by means of semantic tools and resources developed in previous research activities or currently under development. We proposed the identification of the context of features by means of the SemanticNet that could be of relevant interest in order to reach a more complete list of features and attributes related to an object.

Moreover, with the introduction of the synsets and the semantic categorization, instead of

considering only the words as keywords, we aim to define a method of extraction of more accurate meanings and features from textual resources. A validation to support the value of the expressed ideas will be one of the goals of the above mentioned approach and experimental results will be product.

REFERENCES

Akkaya, C., Wiebe, J., Mihalcea, R., 2009. Subjectivity word sense disambiguation. In: *Conference on Empirical Methods in Natural Language Processing*, Singapore, The Association for Computational Linguistics.

Angioni, M., Demontis, R., Tuveri, F., 2008a. A SemanticApproach for Resource Cataloguing and Query Resolution. *Communications of SIWN. Special Issue on Distributed Agent-based Retrieval Tools*.

Angioni, M., Demontis, R., Deriu, M., Tuveri, F., 2008b. SemanticNet: a WordNet-based Tool for the Navigation of Semantic Information. In: *Proceedings of GWC 2008b*. University of Szeged.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M., 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA*. Genoa, Italy. http://nlp.lsi.upc.edu/freeling

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., Subrahmanian, V.S., 2007. Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. In *Proceedings of ICWSM 07 International Conference on Weblogs and Social Media*, pp. 203-206.

Crimson Hexagon, 2009. Listen, Understand, Act. How a listening platform provides actionable insight. http://www.crimsonhexagon.com/PDFs/Crimson_Hexagon_Listen_Understand_Feb_2009.pdf

Ding, X., Liu, B., Yu, P.S., 2008. A Holistic Lexicon-Based Approach to Opinion Mining. *WSDM '08 Proceedings of the international conference on Web search and web data mining*, ACM New York, NY, USA.

Esuli, A. Sebastiani, F., 2007. PageRanking WordNet synsets: An application to Opinion Mining *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* Volume: 45, Issue: June, Publisher: Association for Computational Linguistics, Pages: 424-431

Hatzivassiloglou, V., McKeown, K., 1997. Predicting the Semantic Orientation of Adjectives. In: *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics,* p.174-181, Madrid, Spain

Lee, D., Jeong, O.R., Lee, S., 2008. Opinion Mining of customer feedback data on the web. In: *ICUIMC '08 Proceedings of the 2nd international conference on Ubiquitous information management and communication*

Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., 2002. The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering, special issue on Word Sense Disambiguation*, 8(4), pp. 359-373, Cambridge University Press

Magnini, B., Strapparava, C., 2004. User Modelling for News Web Sites with Word Sense Based Techniques. *User Modeling and User-Adapted Interaction* 14(2), pp. 239-257

Miller, G., 1998. *WordNet: An Electronic Lexical Database*, Bradford Books

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques *CoRR cs.CL/0205070*

Popescu, A.M., Etzioni, O, 2005. Extracting Product Features and Opinions from Reviews. *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing.*(EMNLP'05).

Rentoumi, V., Giannakopoulos, G., 2009. Sentiment analysis of figurative language using a word sense disambiguation approach. In: *International Conference on Recent Advances in Natural Language Processing (RANLP 2009),* Borovets, Bulgaria, The Association for Computational Linguistics

Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, Chun Jin, H., 2007. Red Opal: product-feature scoring from reviews. *ACM Conference on Electronic Commerce 2007: 182-191*

Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM Computing Surveys, 34(1):1–47*.

Turney, P.D., 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *ACL (40th Annual Meeting of the Association for Computational Linguistics)*. Philadelphia, Pennsylvania, USA: ACL.

Vossen, P.(ed), 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W., 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In*: Proceedings of the IEEE International Conference on Data Mining*.

Yi, J., Niblack, W., 2005. Sentiment Mining in WebFountain. In: *Proceeding ICDE '05, 21st International Conference on Data Engineering IEEE Computer Society*. Washington, DC, USA

Zhai, Z., Liu, B., Xu, H., Jia, P., 2010. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, Beijing, China.