AN AUTOMATIC APPROACH TO FEATURE EXTRACTION

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: The pervasive diffusion of social networks as common way to communicate and share information is becoming a valuable resource for analysts and decision makers. Reviews are used every day by common people or by companies who need to make decisions. It is evident that even the opinion monitoring is essential for listening to and taking advantage of the conversations of possible customers in a decision making process. Opinion Mining is a way to analyse opinions related to specific topics: products, services, tourist locations, etc. In this paper we propose an automatic approach to the extraction of feature terms, applying our experience in the semantic analysis of textual resources to Opinion Mining task and performing a contextualisation by means of semantic categorisation, and by a set of qualities associated to the sense expressed by adjectives and adverbs.

1 INTRODUCTION

Social networks are becoming a common way to communicate and to share information, interests, activities and opinions expressed in form of reviews on blogs, forum, or discussion groups. The ability to follow opinions gradually become less adequate and new automatic tools are even more requested and appreciated especially by large organizations that track not only brands but even consumer preferences and opinions. A Gartner analysis for the 2011-year (Gartner, 2011) illustrates the expectations about emerging technologies and how the need for automated methods is growing. Social analytics offers an answer (Crimson Hexagon, 2009), as one of the key themes emerging in the near future.

The main goal of our work is the development of an Opinion Mining system able to extract automatically the features and the meaningful information related and contained in opinions, in a general and not clearly defined domain, from multiple review sources. To achieve this purpose, we propose a linguistic approach to extract and contextualize features related to products or services. We enriched the contents of WordNet, related to the meanings expressed by adjectives and adverbs, with a set of properties having a positive, negative or objective value associated and other properties that could add particular and important information in semantic analysis. The term feature is here used with the same sense given by (Ding et al., 2008) in their approach to Opinion Mining.

As in (Benamara et al., 2007), we propose a linguistic approach to Opinion Mining, based on a combination of adverbs and adjectives. Our work also introduces the use of the WordNet (Miller, 1998) synsets. In order to automatically extract the features from the opinions, the approach is based on the processing of textual resources, on the information extraction and on the evaluation of a semantic orientation. More in details, it performs a syntactic analysis, a semantic disambiguation and a contextualization phase and takes into account the meaning express in conversations.

Moreover the properties related both to adjectives and to adverbs are associated to each synset, providing the result of the analysis grouped into thematic categories.

The remainder of the paper is organized as follows: Section 2 refers to the state of the art and related works. Section 3 introduces the approach and examines the work performed on adjectives and adverbs' structures. Section 4 explains the approach to the feature extraction giving some details about a modified version of the Leacock-Chodorow (Leacock and Chodorow, 1998) algorithm and a demonstration we realized in order to better analyse the results. Finally, Section 5 draws conclusions.

Angioni M. and Tuveri F..
AN AUTOMATIC APPROACH TO FEATURE EXTRACTION.
DOI: 10.5220/0003714504730476
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 473-476
ISBN: 978-989-8425-95-9

2 THE STATE OF THE ART

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". Concerning the task of the development of linguistic resources many works perform WordNet exploring methods and gloss classification methods. In opinion summarization several approaches are based on the use of lexicons of words able to express subjectivity. Akkaya et al. (2009) build and evaluate a supervised system to disambiguate members of a subjectivity lexicon while Rentoumi et al. (2009) propose a methodology for assign a polarity to word senses applying a Word Sense Disambiguation (WSD) process.

Some authors (Lee et al., 2008) asserted that the systems that adopt syntactic analysis techniques on extracting opinion expressions tend to show higher precision and lower recall. One relevant task in opinion summarization step regards the features extraction. Some works about features are based on the identification of nouns through the part of speech tagging (pos-tagging) and provide an evaluation of the frequency of words in the review based on tf-idf criterions and its variation. Others researchers (Zhai et al., 2010) proposed a constrained semi-supervised learning method based on the contextualization of reviews grouped on specific domains and on a characterization based on features defined by users. Other two works are (Popescu and Etzioni, 2005), that worked on the explicit features in noun phrases, and SentiWordNet (Esuli and Sebastiani, 2007) in which WordNet is expanded thanks to a semi-automatic acquisition of the polarity of WordNet terms, evaluating the polarity of each synset.

3 THE APPROACH

Our approach to the feature extraction of terms takes advantage of some tools developed in a previous work on semantic: the Semantic Classifier and the SemanticNet as extension of the WordNet semantic net. Moreover we propose a linguistic approach to extract and contextualize features related to products or services, and a brief description of the new resources we developed: a lexical database of adjectives and adverbs and the semantic distance algorithm.

3.1 A Subjective Lexical Database

Due to the lack of freely available resources having the qualities our approach needs in order to contextualize and group features, a lexical database of synsets has been defined as extension of the properties of WordNet. The resource contains a subset of adjectives and adverbs of WordNet. Each synset has been enriched with a set of properties about the polarity and other properties that could add particular and important information in semantic analysis. We 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. The identified characteristics provide additional information about the content of the sentences, regarding for instance personal, moral, aesthetical aspects or related to geography, to time or to weather for the adjectives. We identified about 15 typologies of adjectives and about 7 typologies of adverbs. Some of these properties allow a polarization that can be used by Opinion Mining algorithms, such as the adverbs of manner.

3.2 The SemanticNet

The SemanticNet has been developed starting from the synsets of the terms contained in WordNet and mapped on the contents included in Wikipedia. It has been defined by adding new nodes, attributes, and new relations named of Common-Sense. defined by means of the conceptual associations defined by the authors in the pages of Wikipedia. Moreover, each page of Wikipedia is associated to the WordNet synset having the most correct meaning by means of a similarity algorithm. In the SemanticNet the nodes are the "senses" (identified by the synset of WordNet or by a unique key defined as "term+category") and links are given both by the WordNet semantic relations and by the conceptual associations built by the Wikipedia authors.

3.3 Feature Extraction

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 it could be very relevant investigate how electorate reacts to politicians statements. Certainly in this case features could not be determined *a priori*, but should be extracted from text automatically. We integrated the approach of (Yi, J. et al., 2003) with some tools in the phase of the extraction of the features, filtering

the list of features through a contextualization phase by means of the semantic classification. In order to extract the features, we essentially make use of syntactic and semantic analysis, the semantic classifier, the subjective lexical database of adjectives and adverbs and the semantic net of WordNet. The syntactic analysis, composed by a pos-tagging and phrase-chunking phase, has been performed in order to identify and extract nouns and, by means of the lexical database, the set of properties related to the adjectives and adverbs.

A distinction between objective and subjective sentences is performed considering the combination of adjectives and adverbs having a polarity associated. Given a collection of reviews related to a specific topic, the semantic classifier categorizes the sentences, extracts as result a set of categories and weights as measures of their relevance in order to define the context for the features in the review. The list of categories with their weights describes the domain referred to the collection of reviews.

For instance, starting from a collection of reviews about tourism and especially linked to hotels, the classifier categorizing the sentences containing opinions produces categories about geographical locations, buildings, staff, food or gastronomy and their weights. These categories are useful in the WSD of the candidate features previously extracted by means of a tf-idf function.

The mapping between synsets related to each feature and the categories, defines a first level of contextualization of the features. The mapping also reduces the list of candidate features to a restricted set having synsets mapped to the top categories. The result is obtained using as reference the categories defined in WordNet Domains (Magnini et al., 2002), a lexical resource, based on the Dewey Decimal Classification system, representing domain associations among the word-senses of WordNet.

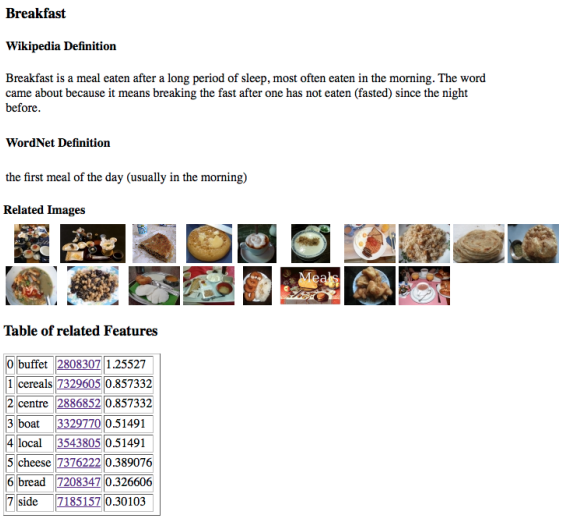
In order to perform the WSD activities a semantic distance algorithm, based on the Leacock-Chodorow algorithm, has been applied. The algorithm calculates the distance measure between feature terms based on information related to the number of synsets referred to each term and according to the mapping of the synsets to the set of categories extracted by the semantic classifier. The algorithm also defines a matrix of the relations of the features having as rows and columns the names of the extracted features and as values the above distance measure. The matrix defines in this way a map of the existing relations between features and, with the introduction of a threshold on the values, puts in evidence different areas of features in order to group them in more specific thematic set.

4 AN INFORMAL TEST CASE

The test is performed on a collection of about 100 reviews of the Alma Hotel of Alghero (Sardinia, Italy) extracted from TripAdvisor, a site of advices and opinions from a huge community of travellers. About 950 sentences have been extracted from the above reviews by the tools for the syntactic analysis.

Tf-idf criterions allows to define an initial set of about 450 candidate features, based on the frequency of their use in the collection.

Once identified, the semantic classifier categorizes the subjective sentences and gives back as result a small set of categories with their associated weights. Then a reduction of candidate features based on the synset mapping on categories has been performed. As final result we obtain a set of about 80 features. As said the semantic distance of synsets and the assignment of weight to the features defines the matrix of the relations between the features extracted. Figure 1 shows the features related to the feature breakfast, one of the features extracted by the system and related to the topic tourism. In the upper side there are two definitions: the first from Wikipedia and the second from WordNet. The SemanticNet in this specific case is



Breakfast

Wikipedia Definition

Breakfast is a meal eaten after a long period of sleep, most often eaten in the morning. The word came about because it means breaking the fast after one has not eaten (fasted) since the night before.

WordNet Definition

the first meal of the day (usually in the morning)

Related Images

Table of related Features

0	buffet	2808307	1.25527
1	cereals	7329605	0.857332
2	centre	2886852	0.857332
3	boat	3329770	0.51491
4	local	3543805	0.51491
5	cheese	7376222	0.389076
6	bread	7208347	0.326606
7	side	7185157	0.30103

Figure 1: The feature breakfast and its related data.

used in order to show the existing mapping between the concepts as defined in the two different resources and a list of related images. Follows the list of the features with the weights measuring their relation to the breakfast feature. Figure 2 shows a

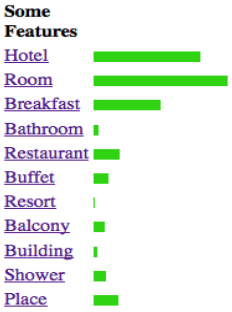


Figure 2: A sample of opinion summarization.

sample of result of the opinion summarization task. The figure illustrates a simple polarity evaluation comparing the preferences about the represented features. In the phase of data aggregation, it is important to notice relations between features, e.g. breakfast and buffet, as showed in the first picture in order to give a correct interpretation of data. The system currently does not analyse this kind of relations between data, but is able to deduce only that users like the breakfast of the hotel but less the buffet organization.

5 FUTURE WORKS

Several Opinion Mining methods and techniques have been developed in order to analyse contents and reviews. In this paper an automatic approach to the extraction of feature terms has been proposed. With the introduction of the synsets and the semantic categorization, we aim to define a method of extraction of more accurate meanings and features from textual resources. We propose the identification of the context of features by means of the semantic net of WordNet in order to reach a more complete list of features and attributes related to an object. Future works will provide the development of the opinion summarization task, the definition of a tool for the navigation of features and related opinions and the generalization of the approach in order to apply it to general contexts. The tool will perform opinion monitoring activities, an essential task in listening to and taking advantage from consumer preferences and opinions. 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.

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*.

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, 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

Gartner, 2011. Gartner's 2011 Hype Cycle Special Report Evaluates the Maturity of 1,900 Technologies. http://www.gartner.com/it/page.jsp?id=1763814

Leacock C. and Chodorow M. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum 1998, pp. 265-283.

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 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

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

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

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*.

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.