e-Review of Tourism Research (eRTR), Vol. 10, No. 2, 2012
http://ertr.tamu.edu

Special Issue – ENTER 2012 Idea Exchange

A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain

Aitor García, Sean Gaines and Maria Teresa Linaza

eTourism and Cultural Heritage Department
Vicomtech-IK4, Spain
[agarcia, sgaines, mtlinaza]@vicomtech.org

Abstract

Sentiment analysis has been extensively investigated during the last years mainly for English language. Currently, existing approaches can be split into two main groups: methods based on the combination of lexical resources and Natural Language Processing (NLP) techniques; and machine learning approaches. This paper introduces the use of lexical databases for Sentiment Analysis of user reviews in Spanish for the accommodation and food and beverage sectors. A global sentiment score has been calculate based on the negative and positive words which appear in the review and using the mentioned lexicon database. The algorithm has been tested with short users online reviews acquired from TripAdvisor.

Keywords: Sentiment Analysis; Opinion Mining; customer review; classification; eTourism.

1 Introduction

Typically, user reviews and opinions allocated in blogs, forums and social networks are short. For instance, the database from TripAdvisor has an average length of 108 words. Furthermore, its style is usually informal and full of different orthographical and grammatical mistakes, misspellings, typos, etc. This poor text quality causes a lot of noise. Therefore, tools designed to analyze formal text suffer a critical performance and accuracy decrease.

This problem becomes harder in languages like Spanish in which words are declined and more difficult to disambiguate in case of misspellings. Also, a lot of heuristics used to analyze English text, like distance between words or their order in the sentence, are slightly more difficult to apply due to the wider variety of word combinations. Our work targets the problem of Sentiment Analysis on the basis of a text included in online accommodation and food and beverage reviews.

Therefore, this paper presents a new approach based on the use of linguistic tools jointly to extract for Sentiment Analysis of online reviews in Spanish. The applied method is based on an annotated lexicon. A global sentiment score has been calculate based on the negative and positive words which appear in the review and using the mentioned lexicon database. The system has been oriented to analyze comments about accommodation and food and beverage.

This paper is organized as follows. Section 2 summaries very briefly some existing approaches in Sentiment Analysis and their application to the tourism sector. An overview of the proposed approach is presented in Section 3. Section 4 presents some details about the evaluation and some conclusions are withdrawn in Section 5.

2 State of the art

2.1 What is Sentiment Analysis?

Sentiment Analysis and Opinion Mining is currently an active research area. It can be defined as the classification of documents based on the overall sentiments expressed by opinion holders (Pang & Lee, 2008). This classification is usually done in positive, negative and (possibly) neutral. Review mining, online reputation, or document summarization are some examples. Their main goal is to extract the global sentiment based on the subjectivity and the linguistic characteristics of the words within an unstructured text.

Several polarity classification methods have been proposed for English. For example, Pang, Lee & Vaithyanathan (2002) proposed a polarity classification method using unigrams and bigrams as input features to several classifiers. The authors used a machine learning approach to exploit the statistical properties of the dataset.

Alternatively, Turney (2002) used the so called Semantic Orientation (SO), which is the Point Mutual Information (PMI) between two words, to calculate the distances to the terms "excellent" and "poor". The output is given by the difference between those two distances ranging between -1 and 1 (negative or positive). Another approach which includes linguistics was introduced in Hu & Liu (2004). They used the WordNet database to find a broad range of positive and negative adjectives based on the distances/hops through the synonyms and antonyms graphs.

One of the closest approaches to this work in Spanish has implemented a linguistic tool to get POS-tagging and to lemmatize words (Moreno, Pineda & Hidalgo, 2010). Authors compare each word of the text with a self-built sentiment lexicon, which contains the context-independent polarity annotation for words with a manifest polarity.

2.2 Application to the tourism sector

Regarding the application of Sentiment Analysis techniques to the tourism sector, Ye, Zhang and Law (2009) analysed the existing approaches to perform automatic classifications based on the Sentiment Analysis of online reviews related to travel destinations, including different supervised machine learning algorithms The algorithms evaluated the reviews about seven popular travel destinations in Europe and North America. Furthermore, Miguens, Baggio and Costa (2008) have analysed a sample of the Lisbon city hotels in TripAdvisor as well as authors and advisors profiles in order to determine the influence of visitors on a destination image.

Finally, the etBlogAnalysis project (Waldhör, 2007) aims at developing a combined crawler /sentiment extraction application for the tourism domain. It combines a simple and robust linguistic parsing methodology with information and terminology extraction methods in order to determine relevant utterances on expression (statement) level. A simple application produces warnings for an enterprise (e.g. hotel) if too many negative entries have been written about it.

3 Overview of the proposed approach

This paper presents a new approach based on the use of linguistic tools jointly to extract for Sentiment Analysis of online reviews in Spanish. The applied method is based on an annotated lexicon which has been self-built and contains more than six thousand sentiment words. A global sentiment score has been calculate based on the negative and positive words which appear in the review and using the mentioned lexicon database. The system has been oriented to analyze comments about accommodation and food and beverage.

Fig1.1 displays the general process overview when analyzing a text and extracting its sentiment. The different blocks indicate the steps followed from the raw text input until the sentiment prediction by category or rating levels is obtained. These steps will be further explained in the following section.

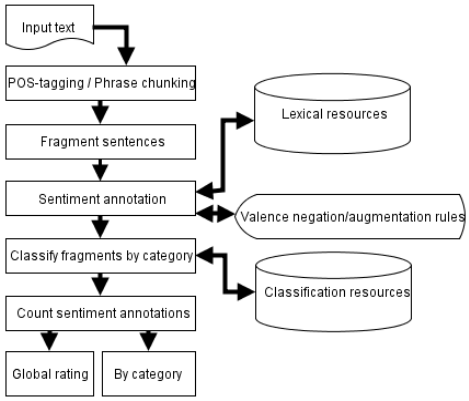

Fig. 1. General process overview

The first step performs a morphological analysis of the text. It also involves term lemmatization in order to have normalized word forms, which is particularly important in languages with words with a lot of declinations. Then, the text is split into fragments, attending to some simple rules, like detecting a noun or noun phrase followed by many complements (adjectives, adverbs, etc.). The intention is to get a list of fragments containing each subject-predicate pair present in the text, using a straightforward and suitable way for this type of informal texts.

The next step is to annotate the obtained fragments according to their polarity. For this purpose, a polarity annotated lexicon developed in our lab has been used. This lexicon contains around 6000 words in Spanish and has been designed for a general domain and context independent polarity. However, the use of the lexicon is not enough to capture the polarity level of main fragments as semantic interpretation should also be taken into account. Therefore, some additional rules so called "contextual valence shifters" (Polanyi & Zaenen, 2006) are mandatory. Rules related to reversion and polarity augmentation have been applied in a straightforward way.

The forth step is related to the sentiment categorization, which can be defined as the extraction of the sentiment for each component of an object. Our approach includes a simple taxonomy to classify fragments by category using a list of lemmatized and normalized words, each of them belonging to a different category or topic. A simple string matching is applied to assign each word with a category. Notice that each fragment can also belong to several categories. Finally, its polarity annotation is counted and aggregated for each categorized fragment.

To obtain the overall rating value, a relative counting criteria of the positive and negative terms has been used. Individual words are assumed to have a prior polarity, that is, a semantic orientation that is independent of context. In such a way, semantic orientation can be expressed as a numerical value. Both positive and negative ratios are in the [0, 1] range, so the overall rating varies in a [-1, 1] range depending on the balance between positives and negatives. The negative sentiments will be closer to -1 and the positive ones to +1, having the neutral ones around the value of 0.

4 Evaluation of the system

Although it is still an ongoing work, preliminary evaluation of the proposed approach has been conducted on the basis of two real datasets of Spanish reviews related to accommodation and food and beverage from TripAdvisor.com. The total number of comments of the dataset are 1000 and 994 for restaurants and hotels respectively. No pre-processing has been performed over the data to deal with misspelling, grammar mistakes, or bad punctuation marks. The obtained accuracy in mere polarity detection has been of 80%

The evaluation has measured the performance for the Sentiment Analysis. We have applied the same sentiment scoring criteria but grouped by categories over a small dataset of 40 human annotated reviews obtaining a sentiment of the different features of an entity. The obtained accuracy in polarity among categories has been of 70.3%.

5 Conclusions

This paper presents a new approach based on the use of linguistic tools jointly with simple classifiers to extract for Sentiment Analysis of online reviews in Spanish. The applied method is based on an annotated lexicon which has been self-built. A global sentiment score has been calculate based on the negative and positive words which appear in the review and using the mentioned lexicon database.

Preliminary evaluation of the proposed approach has been conducted on the basis of two real datasets of Spanish reviews related to accommodation and food and beverage from TripAdvisor.com. Among the preliminary conclusions, it can be mentioned that it seems to be some type of relation between the length of the review and the subjectivity. Therefore, further studies and analysis in other domains will be conducted to clarify the relation between length, number of sentiment words, and subjectivity.

A further conclusion is that negative sentiments are harder to detect than positive ones. Usually, negative sentiments are expressed using an indirect language, irony and also, explaining the whole "negative experience" as a story, which may or may not contain explicit negative words and would require a deeper analysis and world or domain-knowledge to address it.

References

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Miguéns, J., Baggio, R., & Costa, C. (2008). Social media and Tourism Destination: TripAdvisor Case Study. *Proceedings of the IASK Advances in Tourism Research 2008 (ATR2008), Aveiro, Portugal, 26-28 May, 194-199.*

Moreno, A., Pineda, F., & Hidalgo, R. (2010). Análisis de valoraciones de usuario de hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento del lenguaje natural* 45: 31-40.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(1-2): 1-135.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs Up? Sentiment Classification Using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79-86.

Polanyi, L. & Zaenen, A.(2006) Contextual Valence Shifters. In Computing Attitude and Affect in Text: Theory and Applications. Edited by J. Shanahan, Y. Qu, and J. Wiebe. The information Retrieval Series, Vol. 20, Springer, Dordrecht, The Netherlands, pp. 1–9.

Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised classification of Reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Waldhör, K (2007). etBlogAnalysis- Analysing tourism Weblogs and forums using statistical and computer linguistic methods for quality control. *Proceedings of First International Conference on Blogs in Tourism, Kitzbühel, Austria, 12 July 2007*.

Ye, Q., Zhang, Z., & Law., R. (2009). Sentiment classification of online reviews to travel destinations by supervised learning approaches. *Expert Systems with Applications* 36: 6527-6535.

Acknowledgements

This paper is part of the project BEGIRALE- Semiautomatic user-generated comment cataloguing system for the tourism sector, financed by VilauMedia and Alianzo within the GAITEK program of the Department of Industry of the Basque Government.