Opinion Extraction and Classification Based on Semantic Similarities

Aymen Elkhlifi[1], Rihab Bouchlaghem[2] and Rim Faiz[3]
[1]LALIC, Paris-Sorbonne University, 28 rue Serpente Paris 75006, France
Aymen.Elkhlifi@univ-paris4.fr
[2]LARODEC, ISG de Tunis, 2000 Le Bardo, Tunisia
Rihab.Bouchlaghem@isg.rnu.tn
[3]LARODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisia
Rim.Faiz@ihec.rnu.tn

Abstract

This paper presents an automatic extraction and classification approach of opinions in texts. Therefore, we propose a similarity measurement calculating semantically similarities between a word and predefined subgroups of seed words. We have evaluated our approach on the semantic evaluation company "SemEval 2007" corpus, and we obtained promising results: the best value of Precision, 62%; and F1, 61%; as an improvement of 20 % compared to the participant systems.

Introduction

Several techniques have been employed to this purpose like machine learning classifiers, based on lexical features (Bethard and al., 2004), or syntactic features (Wilson and al., 2004) associated to opinion. A wide range of statistical methods are also investigated so as to extract opinions and classify subjective texts (Turney, 2002). In this paper, we contribute to this literature with an alternative strategy of opinions extraction and classification based on semantic similarities between terms.

Our Approach of Opinions Extraction and Classification

Our approach is performed in three steps:

1. **Preprocessing**: consists on the one hand, in the segmentation of text into segments and, on the other hand, in the POS (Part-Of-Speech) tagging.
2. **Opinions-oriented words extraction**: is to extract the opinion-oriented words, by calculating their distances compared to subsets of predefined seed words. We propose in this stage a new similarity measure.
3. **Opinion classification**: consists in classifying the polarity of general opinion based on elementary computing of the second stage.

Opinion-Oriented Words Extraction

We focus in this level on the extraction of all terms holding opinions and the determination of their semantic orientations, which will be used to infer the global opinion valence. Our extraction method is based on a polarity score calculated for every term. So, the word is called subjective if its polarity score exceeds a threshold previously fixed.

To decide which words are opinion-oriented, we propose an algorithm which assigns to each word a score for determination of its polarity. (Bouchlaghem and al., 2010).

Score assignment algorithm: We concede that a term has the same polarity as their direct synonyms. Otherwise, similar words tend to have the same subjectivity class. We adopt these hypotheses and we propose an algorithm by resorting to the similarities between terms and to words synonym sets in order to predict the semantic term orientation. Our strategy is to use a set of predefined seed words. To be able to prepare our seed lists, we undertook an annotation effort of 8000 weak and strong opinion words. Then, we tried to divide this set into subgroups according to subjectivity categories such as: criticism, happiness, harm, approval, joy, etc.

Opinion classification

To predict the polarity of global opinion, we used the average of scores given by our score assignment algorithm. In fact, we calculate a general score based on elementary scores computed in the previous stage, which can also include adjective modifiers such as negation and quantifiers, besides the verbs and nouns. The polarity is classified according to the sign of the overall score: the sentences for which the general score is positive are classified as positive opinions and the sentences with negative scores (Bouchlaghem and al., 2010).

Experimentation and Results

To validate our approach, a system, called Sec-Op (System of extraction and classification of opinions) has been implemented in Java under Eclipse platform. Sec-Op includes the four following modules: Text segmentation, POS tagging, Opinion Oriented terms Extraction and Opinion classification.

We use SEG-EC module (Elkhlifi and al., 2007, 2009) to segment text. Then, we used the Tree tagger API to obtain part-of-speech information. We use WordNet for generating synonyms, and to the measures HSO (Hirst and St-Onge) and LIN for terms similarities computing.

The SemEval 2007 corpus: We have used the SemEval 2007 (the 4th international workshop on Semantic Evaluation) corpus related to Affective text task which is intended as an exploration of the connection between lexical semantics and subjectivity. The corpus consists of 1000 news headlines, extracted from news web sites (such as Google news, CNN) and/or newspapers, and annotated as positive or negative sentences.

Results: To evaluate our approach, we have used the definition of the precision and recall measurements proposed by and modified by (Elkhlifi et al., 2010). Table 1 shows the evaluation results of our system:

TABLE I. EVALUATION RESULTS USING SEMEVAL 2007 CORPUS

	Precision	**Recall**	**F1**
Global performances	62.57	62.06	61.27
Positive class	57.38	73	64.25
Negative class	67.76	51.14	58.28

The method classifies positively better than negativity. Then, we proceeded to compare our results with those of systems participating in the task Affective text of SemEval 2007 (Strapparava and Mihalcea, 2007). The techniques applied by these five systems are various. In fact, UPAR7 is a rule-based system using a linguistic approach. SICS is based on a word-space model and a set of seed words. CLaC is based on a knowledge based domain-independent unsupervised approach. ClaC-NB is a supervised corpus-based system using Machine Learning techniques

We obtained the best value of Precision and F1 (Precision: 62.57 vs. 61.42; F1: 61.27 vs. 42.43). Table 2 shows the interest of our approach.

TABLE II. COMPARISON OF OUR RESULTS (SEC-OP) AND THE RESULTS OF SEMEVAL2007 PARTICIPANTS

	Precision	**Recall**	**F1**
CLaC	61.42	9.20	16.00
UPAR7	57.54	8.78	15.24
SWAT	45.71	3.42	6.36
CLaC-NB	31.18	66.38	42.43
SICS	28.41	60.17	38.60
SEC-OP	**62.57**	**62.06**	**61.27**

Conclusion

The approach proposed comprises three stages to classify the opinion in a text passage, starting, in a first stage, by the preprocessing that consists in segmenting text and tagging its words. In a second step, an algorithm based on P-SIM allows to extract opinion oriented terms. Finally, the polarity of general opinion is predicted with reference to the extracted terms. We validated our approach on a standard corpus from the evaluation company SemEval 2007. The results obtained are promising, comparing to those given by the participating systems.

References

Bethard S., Yu H., Thornton A., Hatzivassiloglou V. and Jurafsky D. 2004. Automatic extraction of opinion propositions and their holders. Proc. The Association for the Advancement of Artificial Intelligence (AAAI-04).

Bouchlaghem R., Elkhlifi A., Faiz, R. 2010. Automatic extraction and classification approach of opinions in texts. In ISDA 2010, Cairo, Egypt. IEEE publisher.

Elkhlifi A. and Faiz, R. 2007. Machine Learning Approach for the Automatic Annotation of the Events. In Proce of FLAIRS 2007. Key West, Florida, USA. pp 362-367.

Elkhlifi A. and Faiz R. Automatic Annotation Approach of Events in News Articles. 2009. In International Journal of Computing & Information Sciences (IJCIS), December 2009, pp 50-60.

Elkhlifi A. and Faiz, R. 2010. French-Written Event Extraction Based on Contextual Exploration. In Proc of FLAIRS 2010, Daytona Beach, Florida. pp 180-185.

Turney P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proc. ACL’02.

Strapparava C. and Mihalcea R. 2007. SemEval-2007 Task 14: Affective Text". Proc. The Fourth International Workshop on Semantic Evaluations (SemEval-2007).

Wilson T., Wiebe J. and Hwa R. (2004). "Just how mad are you? Finding strong and weak opinion clauses". Proc. AAAI’04.