Multilingual Subjectivity and Sentiment Analysis

Rada Mihalcea
University of North Texas
Denton, Tx
rada@cs.unt.edu

Carmen Banea
University of North Texas
Denton, Tx
carmenbanea@my.unt.edu

Janyce Wiebe
University of Pittsburgh
Pittsburgh, Pa
wiebe@cs.pitt.edu

Abstract

Subjectivity and sentiment analysis focuses on the automatic identification of private states, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations in natural language. While subjectivity classification labels text as either subjective or objective, sentiment classification adds an additional level of granularity, by further classifying subjective text as either positive, negative or neutral.

While much of the research work in this area has been applied to English, research on other languages is growing, including Japanese, Chinese, German, Spanish, Romanian. While most of the researchers in the field are familiar with the methods applied on English, few of them have closely looked at the original research carried out in other languages. For example, in languages such as Chinese, researchers have been looking at the ability of characters to carry sentiment information (Ku et al., 2005; Xiang, 2011). In Romanian, due to markers of politeness and additional verbal modes embedded in the language, experiments have hinted that subjectivity detection may be easier to achieve (Banea et al., 2008). These additional sources of information may not be available across all languages, yet, various articles have pointed out that by investigating a synergistic approach for detecting subjectivity and sentiment in multiple languages at the same time, improvements can be achieved not only in other languages, but in English as well. The development and interest in these methods is also highly motivated by the fact that only 27% of Internet users speak English (www.internetworldstats.com/stats.htm, Oct 11, 2011), and that number diminishes further every year, as more people across the globe gain Internet access.

The aim of this tutorial is to familiarize the attendees with the subjectivity and sentiment research carried out on languages other than English in order to enable and promote cross-fertilization. Specifically, we will review work along three main directions. First, we will present methods where the resources and tools have been specifically developed for a given target language. In this category, we will also briefly overview the main methods that have been proposed for English, but which can be easily ported to other languages. Second, we will describe cross-lingual approaches, including several methods that have been proposed to leverage on the resources and tools available in English by using cross-lingual projections. Finally, third, we will show how the expression of opinions and polarity pervades language boundaries, and thus methods that holistically explore multiple languages at the same time can be effectively considered.

References

C. Banea, R. Mihalcea, and J. Wiebe. 2008. A Bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of LREC 2008*, Marrakech, Morocco.

L. W. Ku, T. H. Wu, L. Y. Lee, and H. H. Chen. 2005. Construction of an Evaluation Corpus for Opinion Extraction. In *Proceedings of NTCIR-5*, Tokyo, Japan.

L. Xiang. 2011. Ideogram Based Chinese Sentiment Word Orientation Computation. *Computing Research Repository*, page 4, October.

Tutorial Abstracts of ACL 2012, page 4,
Jeju, Republic of Korea, 8 July 2012.