Sense-level subjectivity in a multilingual setting[☆]

Carmen Banea [a], Rada Mihalcea [a,*], Janyce Wiebe [b]

[a] *Department of Computer Science and Engineering, University of North Texas, Denton, TX, United States*
[b] *Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, United States*

Abstract

Recent research on English word sense subjectivity has shown that the subjective aspect of an entity is a characteristic that is better delineated at the sense level, instead of the traditional word level. In this paper, we seek to explore whether senses aligned across languages exhibit this trait consistently, and if this is the case, we investigate how this property can be leveraged in an automatic fashion. We first conduct a manual annotation study to gauge whether the subjectivity trait of a sense can be robustly transferred across language boundaries. An automatic framework is then introduced that is able to predict subjectivity labeling for unseen senses using either cross-lingual or multilingual training enhanced with bootstrapping. We show that the multilingual model consistently outperforms the cross-lingual one, with an accuracy of over 73% across all iterations.
© 2013 Elsevier Ltd. All rights reserved.

Keywords: Sentiment and text classification; Multilingual subjectivity analysis; Sense level subjectivity

1. Introduction

Sentiment and subjectivity analysis seeks to automatically identify opinions, beliefs, speculations, emotions, sentiments and other *private states* in natural text (Wiebe et al., 2005). Quirk et al. (1985) define a private state as a state that does not lend itself to an objective external validation, or in other words "a person may *be observed to assert* that God exists, but not to *believe* that God exists. Belief is in this sense *private*." (p. 1181). In the field of natural language processing, researchers have used the term subjectivity analysis to denote identifying private states in text, namely separating objective from subjective instances, while sentiment or polarity analysis further refines the subjective text into positive, negative or neutral.

Sentiment and subjectivity analysis has stemmed into a prolific area of research, mainly due to the fact that numerous text processing applications stand to gain from incorporating sentiment dimensions into their models, including automatic expressive text-to-speech synthesis (Alm et al., 1990), tracking sentiment timelines in on-line forums and news (Balog et al., 2006; Lloyd et al., 2005), and mining opinions from product reviews (Hu and Liu, 2004). In many natural language processing tasks, subjectivity and sentiment classification has been used as a first phase filtering to generate more viable data. Research that benefited from this additional layering ranges from question

answering (Yu and Hatzivassiloglou, 2003), to conversation summarization (Carenini et al., 2008), text semantic analysis (Wiebe and Mihalcea, 2006; Esuli and Sebastiani, 2006) and lexical substitution (Su and Markert, 2010).

In experiments carried out on English, Wiebe and Mihalcea (2006) have shown that the most robust subjectivity delineation occurs at sense and not at word level. Following this more fine-grained perspective, Esuli and Sebastiani (2006) and Andreevskaia and Bergler (2006) have proposed methods to embed sense-level automatic sentiment annotations (objective/neutral, negative and positive) over the English WordNet structure (Miller, 1995), using its relationships (synonymy, antonymy, meronymy, etc.). On the other hand, noticing the scarcity of hand crafted sense-level subjectivity/polarity lexica, Markert and Su (2008) have explored ways to infer them from data annotated at either the word or sentence level.

Sense-level subjectivity and cross-lingual subjectivity and sentiment analysis have received considerable attentions in recent years, yet our paper explores the area that lies at the intersection of these two topics. To our knowledge, this area has not been formally investigated, and while the techniques may be similar to those applied in sentiment and subjectivity analysis at the sentence or the review level, our work explores the more difficult task of sense-level subjectivity, which also involves deep semantic aspects of the language. The manual annotation study we performed for this task (cross-lingual sense-level subjectivity annotations), as well as the methods we proposed (cross-lingual and multilingual learning using dictionaries in multiple languages) are novel to our knowledge.

This work seeks to answer the following questions. First, for word senses aligned across languages, is their subjectivity content consistent, or in other words, does a subjective sense in language A map to a subjective sense in language B (and similarly for an objective sense)? Second, can we employ a multilingual framework that can automatically discover new subjective/objective senses starting with a limited amount of annotated data? We seek to answer the first question by conducting a manual annotation study in Section 2. For the second question, we propose two models (see Section 3), one cross-lingual and one multilingual, which are able to simultaneously use information extracted from several languages when making subjectivity sense-level predictions.

2. Sense level subjectivity consistency across languages: annotation study

To answer the first question, we conduct a case study in subjectivity sense transfer across languages, focusing on English and Romanian.

We consider a sense-level aligned multilingual resource such as WordNet. WordNet (Miller, 1995) was first developed for English, and is a lexical resource that maintains semantic relationships between basic units of meaning, or *synsets*. A synset groups together senses of different words that share a very similar meaning. Due to its particular usefulness for NLP tasks, numerous independent non-commercial projects[1] have replicated its structure in over 50 languages, while maintaining alignment with the original WordNet and allowing for sense-level mapping across languages.

In our experiments we use the English (Miller, 1995) and the Romanian (Tufis et al., 2006) versions of WordNet, which contain 117,659[2] and 58,725[3] synsets, respectively. Many of these are aligned at the synset level.

In order to infuse subjectivity information into the model, we use sense-level manually annotated subjectivity data from (Wiebe and Mihalcea, 2006) and (Akkaya et al., 2009), as well as a list of 48 additional words, for a total of 128 words accounting for 580 English senses (with an average polysemy of 4.6). Their equivalent into Romanian is also obtained by traversing the WordNet structure. A native speaker of Romanian (who participated in previous subjectivity annotations studies) was asked to annotate the Romanian data, by being presented with the *gloss* (definition) and the *synset* of each given sense from the Romanian WordNet. The annotator agreement between the English and the Romanian subjectivity labels ranged from 84% (for the Akkaya et al. (2009) dataset) to 90% (for the Wiebe and Mihalcea (2006) dataset). When excluding senses that had both subjective and objective uses in either of the languages, the annotator agreement becomes 87%, with Cohen's $\kappa = 0.74$ for the first dataset, and 94.7% with $\kappa = 0.88$ for the second one, indicating good to very good agreement. These findings support the hypothesis that the subjectivity of a sense maintains itself across language boundaries. Furthermore, they indicate that senses aligned across languages

---