Genre and Domain Dependencies in Sentiment Analysis

Robert Remus

Genre and domain influence an author's style of writing and therefore a text's characteristics. Natural language processing is prone to such variations in *textual characteristics*: it is said to be genre and domain dependent.

This thesis investigates *genre and domain dependencies in sentiment analysis*. Its goal is to support the development of robust sentiment analysis approaches that work well and in a predictable manner under different conditions, i. e. for different genres and domains.

Initially, we show that a prototypical approach to sentiment analysis—viz. a supervised machine learning model based on word n-gram features—performs differently on gold standards that originate from differing genres and domains, but performs similarly on gold standards that originate from resembling genres and domains. We show that these gold standards differ in certain textual characteristics, viz. their *domain complexity*. We find a strong linear relation between our approach's accuracy on a particular gold standard and its domain complexity, which we then use to *estimate our approach's accuracy*.

Subsequently, we use certain textual characteristics—viz. *domain complexity*, *domain similarity*, and *readability*—in a variety of applications. Domain complexity and domain similarity measures are used to determine parameter settings in two tasks. Domain complexity guides us in *model selection* for in-domain polarity classification, viz. in decisions regarding word n-gram model order and word n-gram feature selection. Domain complexity and domain similarity guide us in *domain adaptation*. We propose a novel domain adaptation scheme and apply it to cross-domain polarity classification in semi- and unsupervised domain adaptation scenarios. Readability is used for *feature engineering*. We propose to adopt readability gradings, readability indicators as well as word and syntax distributions as features for subjectivity classification.

Moreover, we generalize a framework for *modeling and representing negation* in machine learning-based sentiment analysis. This framework is applied to in-domain and cross-domain polarity classification. We investigate the relation between implicit and explicit negation modeling, the influence of negation scope detection methods, and the efficiency of the framework in different domains. Finally, we carry out a *case study* in which we transfer the core methods of our thesis—viz. domain complexity-based accuracy estimation, domain complexity-based model selection, and negation modeling—to a gold standard that originates from a genre and domain hitherto not used in this thesis.