Project Proposal for cs228

Chinmay Kulkarni and Gabor Angeli

1 Problem Statement

2 Previous Work

2.1 Automatically Assessing Review Helpfulness (Kim et al., 2006)

This paper was the earliest work present in the literature on the subject of assessing review helpfulness. The main idea of the paper was to use SVN regression on a large number of features, analyzing the results and the impact of each subset of features on the performance of the system. The paper concluded that a relatively simple feature set (length of the review, unigram counts, and star rating) performed the best, with the first two (length and unigram counts) alone being almost as effective as all three combined.

The dataset used by the paper was a corpus of Amazon reviews, in the categories of *MP3 Players* and *Digital Cameras*. The dataset consisted of 821 products and 33,016 reviews for *MP3 Players*, and 1,104 products and 26,189 reviews for *Digital Cameras*. This data was then filtered for duplicate or near-duplicate entries ($> 80\%$ bigrams match), resulting in 85 products and 12,07 reviews being discarded for *MP3 Players*, and 38 products and 3,692 reviews being discarded for *Digital Cameras*. Helpfulness ratings were obtained by taking every review with more than 5 helpfulness responses. This resulted in approximately a third to half of the reviews being discarded.

Training was done using SVM regression. The authors tested on a variety of kernels, however found a radial bias function (RBF) kernel to perform the best; all results are reported using this method.

To assist in feature creation, a *Product-Feature* set was automatically extracted. This was done by mining references to product features from `Epinions.com`, where users are allowed to describe the pros and cons of each product. Frequent words were pruned from this list, resulting in around 10,000 unique features for both domains.

Features were extracted over each review. These features fell into the following categories:

1. **Structural**: review length (`LEN`); average sentence length, number of sentences, etc. (`SEN`); HTML formatting (`HTM`)
2. **Lexical**: *tf-idf* statistic on each unigram (`UGR`); *tf-idf* on each bigram (`BGR`)
3. **Syntactic** (`SYN`): features on POS information
4. **Semantic**: product features (`PRF`); *General Inquirer* sentiment words (`GIW`)
5. **Meta Info**: star rating (average/deviation) (`STR`)

Evaluation was done using the Spearman correlation coefficient. The best results came from the three features (`LEN+UGR+STR`), resulting in a Spearman coefficient of 0.656 on *MP3 Players* and 0.595 on *Digital Cameras*. Adding additional features tended to hurt performance;

adding every feature dropped the *MP3 Player* score to 0.601, although mildly improving the *Digital Camera* score to 0.604.

2.2 A Joint Model of Text and Aspect Ratings for Sentiment Summarization (Titov and McDonald, 2008)

This paper presents an approach to summarization, jointly learning the topics to summarize and text which describes them. The paper proposes a model (*Multi-Aspect Sentiment Model*) consisting of two parts: an unsupervised topic model, and a classifier from words to sentiment ratings. The paper evaluates on a hotel review dataset taken from `TripAdvisor.com`.

The paper presents the task of summarization as a two-fold task: the first task is described as *aspect identification and mention extraction* – determining which aspects of the reviews are relevant to describe, and determining which text fragments describe them. The second task is *sentiment classification* – determining the sentiment on the relevant extracted text. The paper attempts to incorporate both of these tasks into a single model which extracts text fragments and their associated rating, given the review and a per-aspect rating.

The dataset used in the paper consists of 10,000 reviews from `TripAdvisor.com`, where each review was rated in at least *service*, *location*, and *rooms*.

The approach taken is the build a model – coined as a Multi-Aspect Sentiment model (MAS) – which is effectively built on a combination of a multi-grain LDA topic model and a series of MaxEnt classifiers for each topic. The model is such that a word in the document is sampled from either a local or global topic; the intent is that global topics will capture topics corresponding to non-sentiment phenomena (e.g. *MP3 players* versus *hotels*), while the local topics will capture sentiment-laden words.

The paper raises the issue that often aspects of reviews correlate strongly with each other. That is, if you dislike a hotel, you will likely not rate any aspect highly. To address this, the model classifies not over absolute ratings, but over the difference between the aspect rating and the overall rating.

Inference over the model was done using Gibbs sampling, as exact inference is intractable. The model achieves a precision of between 75% and 85% for the different aspects.

References

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430, Sydney, Australia, July. Association for Computational Linguistics.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, June. Association for Computational Linguistics.