Mining Opinion Features in Restaurant Reviews

Xilian Li, Wenjia Ma, Wen Cui

Oct 28, 2017

1 Introduction

Everyone has the experience that when you search restaurants around you on websites like Yelp or TripAdvisor, you will not only check the rating but also the reviews to get an idea of the quality of food and service especially when you plan an important event in that restaurant. The number of customer reviews grow rapidly as these websites are becoming popular. In this project, we aim to automatically summarize the potential attributes of restaurants in a more concise way, that is mined from learning the reviews. So that without reading through hundred of reviews, customer can have an insight of the restaurant's features and rating of each features.

2 Related work

Many researchers have contributed to analysis of product reviews and laid a foundation for our work. Hu and Liu[5] find useful patters to extract features by applying class sequential rule mining. . Chen et al.[6] proposed that domain knowledge can largely improve the outcome of feature extraction. Wang and Ren [2] point out a way to calculate the sentiment score of each extracted frequent feature. While much research are on product reviews, such as electronic products, we want to test the algorithms and generate feature scores on reviews of restaurant, which contain more features than product(food) itself, such as the environment and the experience, etc. There surely will be more researches to be explored and include when we start to work on the project.

3 Dataset

We'll use reviews data from Yelp dataset which is available here[1]. This dataset contains all kinds of reviews of local business. So we need to extract just restaurant reviews from the dataset by identifying the business_id and determining which business_id are restaurants. We don't know yet exactly how many restaurant reviews after the seperation but it must be more than thousands.

4 System Overview

4.1 Preprocessing

Firstly we will preprocess the data by cleaning non-textual content and do some statistic analysis on data distribution. For example the distribution of review length and ratings. After we get to know the data, we'll apply natual langugae processing pipline like tokenization, POS tagging and dependency parsing.

4.2 Feature Extraction

After the preparation above, we should be able to extract features which is (attribute, opinion) pair by the dependency of sentences. For example

The food is great. → amod(food, great)

This indicates amod is a good pattern of feature. In this stage, we need collect some patterns as candidate feature sets.

4.3 Feature Pruning

Frequent Feature Generation. This step is to find features that people are most interested in and talk about the most. And we will use Apriori algorithm [3] which finds all frequent itemsets in the candidate set. Each resulting frequent itemset is a possible feature. In our work, we define an itemset as frequent if it appears in more than 5% (attempted minimum support) of the review sentences.

Semantic Pruning. Some frequent features generated in the first step may not be as useful or may overlap with others. So in this step we focus on the semantic meaning of the features. We will use the semantic distance between the feature words and the restaurant by either applying Wordnet or Word2Vec language model. And we only select the features that the distance is more than certain threshold (0.6 for example).

4.4 Sentiment Detection

In this step, we will apply sentiment score using sentiment analysis on each adjective word in the feature set generated above. And for each attribute, we will have a number of words describing it. By summing up the sentiment scores range from -1 to 1, and normalize it to get the overall score for this attribute. For example, we may end up with *'food'* attirbute has opinion words $\{good, excellent, fresh\}$, so that we can calculate the score as follows,

$$\begin{cases} good = 0.5 \\ excellent = 0.8 \\ fresh = 0.3 \end{cases} \Rightarrow food = \frac{0.5 + 0.8 + 0.3}{3} = 0.53$$

Note the score below is in [-1, 1] scale, which we can easily scale up to map 0-5 rating scale.

4.5 Summary

After the process above, our system automatically output for each restaurant a summary like this, all the rating are in 0-5 scale as the convention of Yelp.

RESTAURANT NAME: *Woodfired Pizza*
PROS: *price(4.2), fresh(4.5)*
CONS: *waiting(2.1)*

5 Evaluation

Due to the data and time limitation. We will evaluate our feature selection by manually annotation of a subset of data. And calculate the accuracy, precision, recall and f-measure of our system defined below,

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F - measure &= 2 \cdot \frac{precision \times recall}{precision + recall} \end{aligned}$$

6 Tools

- WEKA for cleaning and mining https://www.cs.waikato.ac.nz/ml/weka/
- python NLTK for text processing http://www.nltk.org/book/
- Stanford Core NLP https://stanfordnlp.github.io/CoreNLP/
- Google cloud NLP https://cloud.google.com/natural-language/

7 Schedule

Oct 28(Sat):Complete Proposal Nov 4(Sat): Complete Preprocessing and visualization of data Nov 11(Sat): Complete Feature Extraction and feature pruning Nov 18(Sat): Complete Sentiment Detection and evaluation Nov 25(Sat): Evaluate, analysis, and modification Dec 2(Sat): Revise project report Dec 5/8: Paper presentation

References

[1] Yelp dataset https://www.yelp.com/dataset/challenge

[2] Jingye Wang, Heng Ren *Feature-based Customer Review Mining* `https://nlp.stanford.edu/courses/cs224n/2007/fp/johnnyw-hengren.pdf`

[3] R. Agrawal, R. Srikant. "Fast algorithms for mining association rules in large databases" *Proc.of 20th Int'l conf. on VLDB: 487-499, 1994.*

[4] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis , Jeff Reynar, Building a Sentiment Summarizer for Local Service Reviews

[5] Minqing Hu, Bing. Liu, Opinion feature extraction using class sequential rules, Paper presented at the AAAI Spr

[6] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, Riddhiman Ghosh, Exploiting domain knowledge in aspect extraction, Paper presented at the Emnlp (2013)