SemSort: Sentiment Analysis Sorting

Summarizing and Sorting Yelp Reviews

Wenhao Zhang, Graeme Milne, Mitchell McCormack & Jonathan Lo
Simon Fraser University
Group: Wisefish, CMPT 413

WiseFish

Introduction

SemSort utilizes two natural language processing artificial intelligence techniques to automatically aggregate reviews for given business into quickly digestible lists. Review summarization is done using an improved SumBasic algorithm, a multidocument summarization tool. Semantic Analysis is carried out by a neural network trained to predict the positive or negative sentiment of sentence. SemSort is a tool to provide additional subjective information to reviews, offering an extension to Yelps star rating system.

SemSort Pipeline

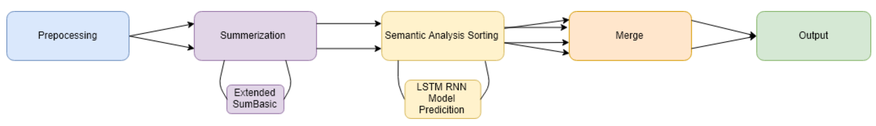

Figure 1: Pipeline for Processing Yelp Reviews

1. Preprocess the data to prune unnecessary information
2. Summarize Yelp reviews for a business
3. Sort the summarized sentences into positive and negative lists
4. Merge the data streams into a single coherent output

Preprocessing

We are using the Yelp Data Set [5] as our primary source of data. This dataset contains a lot of unnecessary data as our pipeline only needs the business location, the reviews it has received, and the average star rating of these reviews.

To trim the data that the later steps in the pipeline, we perform the following preprocessing steps:
Step 1 Load all from the reviews from the 4GB dataset.
Step 2 Filter to only include reviews for businesses in Toronto, Ontario, Canada. Additionally, keep only businesses that have an adequate number of reviews for evaluation.
Step 3 Keep only the necessary data fields: Business ID, Review Count, Stars, and the Reviews.

SumBasic

SumBasic is a summerization system for multi document input that uses word probability to determine the importance of sentences. Each sentence is assigned a weight equal to the average probability of the words in it. Then the best scoring sentence is selected. The algorithm runs until the generated summary meets the expected length. The full steps to the SumBasic algorithm are shown below.

Step 1 Compute the probability distribution over the words w_i appearing in the input, $p(w_i)$ for every i; $p(w_i) = \frac{n}{N}$, where n is the number of times the word appeared in the input, and n is the total number of content word tokens in the input.
Step 2 For each sentence SJ in the input, assign a weight equal to the average probability of the words in the sentence, i.e.

$$weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i}{|\{w_i | w_i \in S_j\}|}$$

Step 3 Pick the best scoring sentence that contains the highest probability word.
Step 4 For each word w_i in the sentence chosen at step 3, update their probability

$$p_{new}(w_i) = p_{old}(w_i) * p_{old}(w_i)$$

Step 5 If the desired summary length has not been reached, go back to Step 2.[4]

Opinosis

Opinosis is an abstractive summarization method that makes use of graphs to generate summaries of highly redundant opinions. Unlike many other methods Opinosis does not require any domain knowledge and only uses shallow NLP processing.

Validation

The validation of the SumBasic implementation in this project was conducted using the Opinosis data set using ROUGE-N for the evaluation metrics. The ROGUE-N metrics measure the N-gram overlap between the SumBasic generated summary and one or more of the gold standard reference summaries in the test data set. For the purposes of this project, ROGUE-1 and ROGUE-2 precision and recall were used.

	ROGUE-1	ROGUE-2	Avg Num Words
SumBasic	0.3037	0.0777	34
Graph-Based Opinosis Paper[1]	0.2831	0.0853	15

Table 1: Recall

	ROGUE-1	ROGUE-2	Avg Num Words
SumBasic	0.1538	0.0377	34
Graph-Based Opinosis Paper[1]	0.4482	0.1416	15

Table 2: Precision

Semantic Sorting

Semantic sorting is carried out by passing each sentence returned from the extended SumBasic algorithm into a model to predict if the sentence is positive or negative. The model was created using a Long Short Term Memory Recursive Neural Network[3]. The training data was preprocessed by sorting a sample of yelp reviews into positive and negative categories based the star rating the review had. One and Two stars were considered a negative review. Four and five stars were considered a positive review. Then Three methods were tested for training the model.

The first method was to pass the entire text of the review through the network with it's label of 'Positive' or 'Negative'. The second method was to split the text of the review into sentences and pass the sentences through the network individually with the corresponding label. The third method experimented with, passing all of the bigrams generated from each sentence as a bag of words. Method 3 was undertaken as an adaptation of FastText[2] to a RNN and as the results show was the least effective.

The sorting method is to take the sentences produced by the summarizing algorithm then pass each of them into a model to produce two values. The sentence is sorted by the argmax of the values produced by the prediction. Model 1 is used to produce sentiment predictions. Table 3 shows the considerable increase in accuracy Model 1 has over Models 2 and 3.

Results

Summarization

The implementations of SumBasic and Opinosis have similiar results. The impementation of the graph-based summarizer does not match the results of the paper but is relatively close.

Semantic Analysis

Method	Positive	Negative
Model 1	0.944	0.939
Model 2	0.770	0.777
Model 3	0.584	0.833

Table 3: Model Prediction Accuracy

In training a RNN, the format of the input data is a key component in determining the accuracy of the model. The best result observed from experimenting with different formats for the input data was to pass the entire review text with it's corresponding label. This approach yielded a 17.4% increase in accuracy over a individual sentence approach and a 36.0% improvement over passing labeled bigrams to the network.

SemSort

Output from the 252 reviews of Toronto's Eaton's Center

Positive:

- "The Eaton Centre does offer some positve exceptions, even considering the national demise of its namesake several years ago.
- The Hudson Bay store on one end is a fine department store that wears its age and history well.
- And when you consider all the other activities that Toronto has to offer within walking and transit distance, the Eaton Centre is just another interesting and fun choice within this vibrant metropolitan area.
- Most interesting is Trinity Square which is literally just outside the door near Sears and on the opposite side from Yonge.

Negative:

- Even at the Mall of America, the West Edmonton Mall, and this modernly designed and spacious Eaton Centre, the shopping experience soon becomes similar to any other big mall.

References

[1] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, pages 340–348. Association for Computational Linguistics, 2010.

[2] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759, 2016.

[3] Peter Nagy. Lstm sentiment analysis — keras, Feb 2017.

[4] Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization.

[5] Yelp. Yelp open dataset, 2018.