ONLINE SHOPPING REVIEW SUMMARIZATION FOR PEOPLE WITH VISUAL IMPAIRMENTS

12 – 02 – 2019

Yinchen Xu, Qiwen Wang, Zhixiu Liu

INTRODUCTION

Blind and visually impaired(BVI) people often struggle to find out information on products when shopping online. Online reviews provide a promising source, yet remains infeasible for BVI users to search through a large amount of text. Hence, we propose a review summarization pipeline in the food category with a novel approach that deploys transfer learning on the BERT model with both extractive and abstractive approaches. The extractive summarization approach chooses the most relevant sentence in the review, while the abstractive summarization approach generates the summary word by word. We evaluate models using the precision, recall, F1 score, and Rouge-n score, a popular metric for the text summarization task that computes the N-gram co-occurrence statistics.

METHODS

Our task is to generate a concise, accurate and fluent summary based on all reviews of an item.

Evaluation Metric: We used F1 score, F1-score of ROUGE-2, and precision to evaluate our summaries.

Data: We used Amazon Fine Food Review which consists of ~500,000 reviews and summaries up to Oct 2012 from all kinds of its categories. We used CNN/DailyMail for training.

Word Embeddings: We used summation of pretrained BERT embeddings, positional embeddings as well as segmentation embeddings.

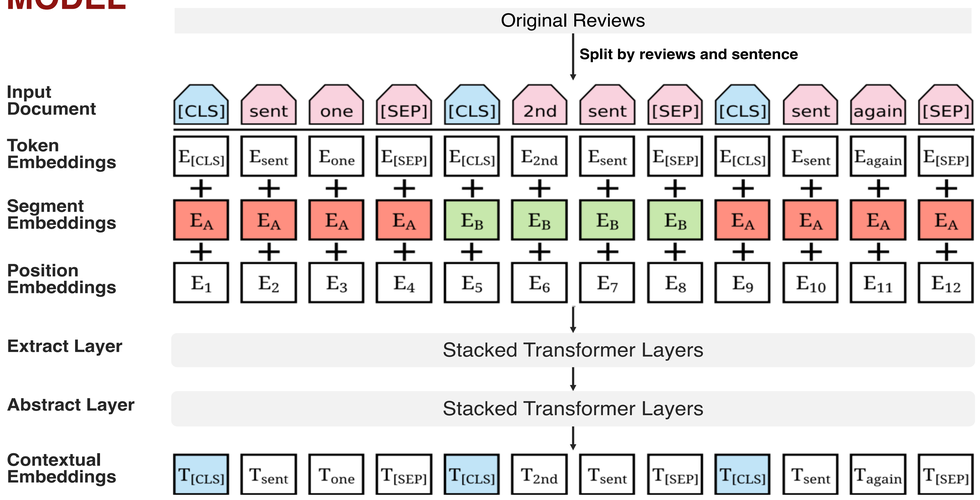
Preprocessing: 1. Dropped duplicates and NA values. 2. Grouped the reviews by the product id and select the data from the frequently reviewed product. 3. Expanded the contracted words such as 'I'm' and 'there'll'. 4. Removed the stop words. 5. Split each review into words.

Training: For the extractive model, the top 3 sentences are selected based on the rankings of the reviews. The abstractive model is built upon the trained extractive model and words are consecutively generated until hitting the end label. Extractive & Abstractive models have independent Adam optimizers to separate the pre-trained parts and the abstractive parts trained from the scratch. We use dropout probability 0.1 before every linear layer.

RESULTS

Model	Precision	Rouge-2 Score	F-1 Score
Baseline: K-means	7.04	16.7	12.34
Human-generated oracle	4.01	3.57	5.81
LSTM w/ self-attention	36.36	N/A	24.24
Bi-LSTM w/ self-attention	36.36	N/A	24.24
Bert w/ extractive	52.62	7.43	54.67
Bert w/ extractive & abstractive	51.37	3.16	52.47

MODEL



We first used Stanford CoreNLP to tokenize and split the raw review inputs. Then we applied transfer learning based on the BERT model, a popular text representation model for NLP tasks. The model uses a two-stage fine-tuning approach, where the BERT model uses the weight trained with an extractive objective that evaluates the relevance of each review sentence to train on the abstractive method, which generates the sentence word by word. The encoder and decoder are inter-sentence transformers representing the multi-sentence relevance of the same product.

DISCUSSION

- Quantitative evaluation methods for the generated summary can be efficient in testing certain aspects of the summary, but each of them has certain restriction. Precision can have bias to shorter summary. ROUGE-2 only concentrates on the logic in spatial locality in the generated result. Combining multiple metrics and user evaluations are encouraged for comprehensive evaluations.
- As observed by several users, though BERT model with extractive and abstractive layers ranks lower in quantitative metrics than model only with extractive layer, the generated summaries read more concise and conclusive.
- Similarity with original reviews is an obvious weakness for methods based on BERT models. We expected a better performance in this aspect for models with abstractive layers, but since the training pipeline requires that abstractive layers are trained based on the fine-tuned weights of the extractive layers, the difference is minimal.

CHALLENGES

- As the extractive model merely pick relevant contents from the original reviews, the results may not seem comprehensive and natural. Meanwhile, although the abstractive model might tackle this problem by generating results word-by-word, its results are similar with those of the extractive model.
- From the perspective of generating meaningful summarizations, the length of the summarized results should be proportional to the original length of the reviews. However, all of the models we implement are of fixed length, which would result in the loss of contents when it comes to longer original reviews.
- The reviews of our oracle are mostly words phrases however, the data we trained on is mostly sentences, which resulted in a deviation from our oracle.

Stanford University

Computer Science Department
Stanford University, Stanford, California 94305
cs.stanford.edu