Extractive Summarization on Yelp Reviews

Kekoa Riggin
University of Washington
kekoar@uw.edu

Abstract

The results of creating a tf-idf-based extractive summarization system for the Yelp Dataset Challenge Round 12. Using reviews of the same star rating as a docset, summarization can extract representative sentences from reviews to demonstrate what businesses are doing well and not so well.

Introduction

Summarization is an NLP task that has many practical applications in the data-driven age. The goal of summarization is to identify the core concepts of a large text and present those concepts with a low cognitive cost.

Yelp reviews are a useful tool for individuals to share and access opinions regarding business. The catch 22 of Yelp reviews is that the confidence of reviews for a particular business increases directly with the number of reviews written, but the more reviews there are, the less practical it becomes to read them all.

Summarization provides a solution to accessing the value of reviews without the cognitive cost to the user of reading a large body of text with varying writing styles and authors.

This paper documents the features and performance of a summarization system designed to solve this problem.

System

This system generates summaries of a collective set of Yelp reviews based on their star rating.

The features and methods for summarization are detailed in the following subsections.

Extractive Summarization

The summarization engine used in this project is an extractive system. Extractive summarization uses segments that are present in the system to compile a condensed representation of the entire body and is the counterpart to abstractive summarization, which generates original summary text (Extractive Summarization).

Modified tf-idf

In order to select segments that contribute to an informative summary, this system uses a modified version of the tf-idf weighting scale (tf-idf).

In this system, each word found in a review is treated as a term, or as Conroy et al. put it, an idea (2004). These terms comprise the message of the review, and the goal of this system is to identify the central idea of the reviews. A tf-idf score is a way of evaluating the centrality of each term.

Traditional tf-idf scores, however, favor terms that have a high density in few documents (Riggin et al. 2018). In order to find the central message of the reviews of a single star rating, this system uses a scoring metric that favors terms that appear in many documents in the target star rating but reduces the score of a term that appears often in other star ratings as well.

The weighting metric used in this system is a tf-df*itf-idf, where tf-df is the term frequency multipled by document frequency in reviews of the target star rating and that is multiplied by itf-idf, which is the inverse term frequency multiplied by the inverse document frequency in reviews belonging to reviews with star ratings outside of the target.

Data

The data comes from the reviews file from the Yelp Dataset for Round 12.

The data used for summarization is extracted from the "text" value of each review.

In modeling the data, each review is considered as a document. All reviews from one business that have the same star rating comprise a docset. Thus, each business has 5 docsets (one for each star rating), and each docset is the collection of reviews with that star rating.

Businesses used in this experiment were selected for having a similar number of reviews per star rating and a large enough number of reviews to generate extractive summaries. This system could run on any business in the dataset, but its reliability on smaller docsets may vary.

Of the businesses used, there are:

- 1 Business with 100-150 reviews per star rating
- 8 Businesses with 100-200 reviews per star rating
- 2 Businesses with 150-250 reviews per star rating
- 30 Businesses with 50-99 reviews per star rating

Evaluation

The best method for assessing the quality of summarization requires manually curated data. Quality metrics such as ROUGE or BLEU use human-generated summaries as a baseline and compare the system output to those summaries (ROUGE and BLEU).

Unfortunately, this project doesn't have the resources for completely curated evaluation.

However, the evaluation of one generated summary for one business has been manually and included in this report.

All other evaluation has been completed using a token coverage script that evaluates the total percentage of tokens that are covered by the generated summaries.

This script evaluates coverage based on the n-grams up to 3-grams that are present in the reviews of a star rating and the generated summary for that star rating. These n-gram exclude any grams with punctuation or tokens from a stop list from Yoast SEO (Yoast).

The coverage score is the percentage of grams covered by the summary over the total number of grams in the reviews. Because the model from Conroy et al. considered each term as an idea, a general coverage score theoretically demonstrate the total ideas captured by the generated summary. The more ideas covered, the better representation of the reviews as a whole is captured in a snapshot of the text.

The qualitative value of terms is not measured by term coverage. This measurement indicates that the system consistently captures a wider range of terms (or ideas) than randomly generated summaries. Additional evaluation is necessary to determine whether the terms capture the central idea of reviews.

Performance

For a lack of resources, performance was measured manually for only one business. Additionally, an evaluation script was used for simple token coverage analysis.

Manual Analysis

System performance was evaluated manually for the business with id: **VUtazCTIc0aoOrQprP_s-Q**.

This evaluation was completed by reviewing the text of each review and tallying the number of opinion-based mentions of anything about the business. This included things like service, wait time, menu items, music, various things about the location, etc. The tallies were then normalized by the percent of mentions.

For each of the five generated extractive reviews, 10 reviews were generated by randomly selecting sentences from the entire body of reviews until the random review was larger than the extractive review. These random reviews were used as a baseline for comparison with the generated review.

For each evaluated review, every mention from the tallies added the normalized tally count to the review's score. This score shows how often central ideas from the reviews appear in the summary.

The results are seen in the table below.

Extractive	Random Average
29.2	22.82

Based on the results of this evaluation, the generated extractive summary contains sentences that represent central ideas with greater frequency than randomly generated summaries.

Coverage Analysis

Analysis of token coverage was performed on two sets of data. The first set is a group of 10 businesses that have a range of 100-250 reviews per star rating. The second is a group of 30 businesses that have a range of 50-99 reviews per star rating.

For each generated extractive review, 10 random reviews were generated. The scores for these random reviews were calculated as averaged to be used as a baseline against the extractive system.

The results are as seen in the table below.

	Set 1	Set 2
Average Extractive Coverage	3.3%	6.0%
Average Random Coverage	2.9%	5.3%
Average Difference	+0.6%	+0.7%
Difference (tokens)	+10.3	+15.6

Coverage analysis shows that the extractive system consistently covers a larger portion of the total tokens in both sets.

Discussion

This system generates viable summaries of Yelp reviews that present core concepts found in the original review text. Based on manual evaluation and token coverage, the system consistently performs better than randomly generated summaries of equivalent or greater length.

In addition to selecting sentences with core concepts, the sentences that appear in generated extractive summaries are typically grammatical, while randomly generated sentences had a higher rate of ungrammatical or unintelligible sentences.

This system doesn't only produce summaries that contain central concepts with greater frequency, but also covers a greater quantity of the concepts contained in the original reviews.

The results from token coverage analysis indicate that the generated extractive reviews cover between 10-15 more tokens than randomly generated summaries.

Given that the sentences average 15 tokens per sentence, the token discrepancy is nearly the value of an entire sentence. This occurring while randomly generated sentences have a greater average length than extractive reviews.

The results of this project show promise that summarization could provide value to end users of business review apps like Yelp. Summaries reduce the cognitive cost and quantity of text that a user must read in order to access the core concepts contained in a body of reviews.

Further Study

In continuing the work in this project, further qualitative analysis of the system's performance is the first step. The coverage evaluation, although a reliable measurement of the ideas captured by the generated summary, does not indicate the holistic value of the summary.

Manual analysis allows more additional measurements to be made concerning summary quality, and also makes using standard summarization quality metrics such as ROUGE or BLEU possible.

In addition to further evaluation of this system, other summarization methods are worthy of investigation.

In a future study, this system could be improved by adding features that improve content selection or realization. Additionally, abstractive methods for summarization should be explored as a solution to the review summarization question.

Access the project at https://github.com/unclenachoduh/yelp12 (https://github.com/unclenachoduh/yelp12)

Sources

- Conroy, J. M., J. D. Schlesinger, J. Goldstein, and D. P. O'Leary (2004) Left Brain Right Brain Multi-Document Summarization.
- Riggin, Greve, Lindberg, Mathias, Topic-focused Summarization via Decomposition Deliverable D3. Unpublished, Ling573 University of Washington. 2018
- BLEU Scoring Metric https://en.wikipedia.org/wiki/BLEU (https://en.wikipedia.org/wiki/BLEU)
- Extractive Summarization https://en.wikipedia.org/wiki/Automatic_summarization#Extraction-based_summarization (https://en.wikipedia.org/wiki/Automatic_summarization#Extraction-based_summarization)
- ROUGE Scoring Metric https://en.wikipedia.org/wiki/ROUGE_(metric (https://en.wikipedia.org/wiki/ROUGE_(metric))
- tf-idf https://en.wikipedia.org/wiki/Tf%E2%80%93idf (https://en.wikipedia.org/wiki/Tf%E2%80%93idf)
- Yoast Wordpress SEO https://github.com/Yoast/wordpress-seo (https://github.com/Yoast/wordpress-seo)