Extraction-Based Summarization of Restaurant User Reviews

Venkata Sarika Kondra
C00219805@louisiana.edu

Agenda

- Text Summarization
- Latent Semantic Analysis
- Proposal 1 : Summarizing user reviews using LSA
- Proposal 2 : Multilabel classification of user reviews into relevant categories
- Proposal 3 : Evaluation of text summarization results
- Resources
- References

Text Summarization

Due to the overloaded content available on the web, there is a serious necessity for text summarization tools for many applications and users. Text summarization solves the problem of presenting the information needed by a user in a compact form.

History - First summarization work started in 1950s [2]

Applications of summarization - document summarization, image summarization, video summarization, reviews summarization etc.

Types of Summarization:

1. Extraction-Based Summarization
2. Abstraction-Based Summarization

Review Summarization for restaurants

It is becoming increasingly difficult to handle the large number of opinions posted on review platforms and at the same time offer this information in a useful way to each user so he or she can make a decision fast enough in visiting a restaurant or not.

Topic-based aggregations and short review summaries are used to group and condense long user reviews to shorten the decision time of a new customer.

Proposal 1: Summarizing user reviews using LSA

Latent Semantic Analysis-

It is an algebraic-statistical method that extracts hidden semantic structures of words and sentences. It is an unsupervised approach that does not need any training or external knowledge.

Singular Value Decomposition is used to find out the interrelations between sentences and words.

Latent Semantic Analysis

- Create a term by sentences matrix A = [A1, A2, …, An] ,

  where,

  - Ai represents the weighted term-frequency vector of sentence i in the document under consideration.
  - If total terms = m and total documents = n, then A is a matrix with dimensions m * n

Singular Value Decomposition

- A is a very sparse matrix, performing SVD on A results in,

$$\mathbf{A = U\Sigma V\ ,}$$

  - where U = [uij] is an m × n column-orthonormal matrix whose columns are called left singular vectors;
  - Σ = diag(σ1, σ2, …, σn) is an n × n diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order, and
  - V = [vij] is an n × n orthonormal matrix, whose columns are called right singular vectors

Singular Value Decomposition

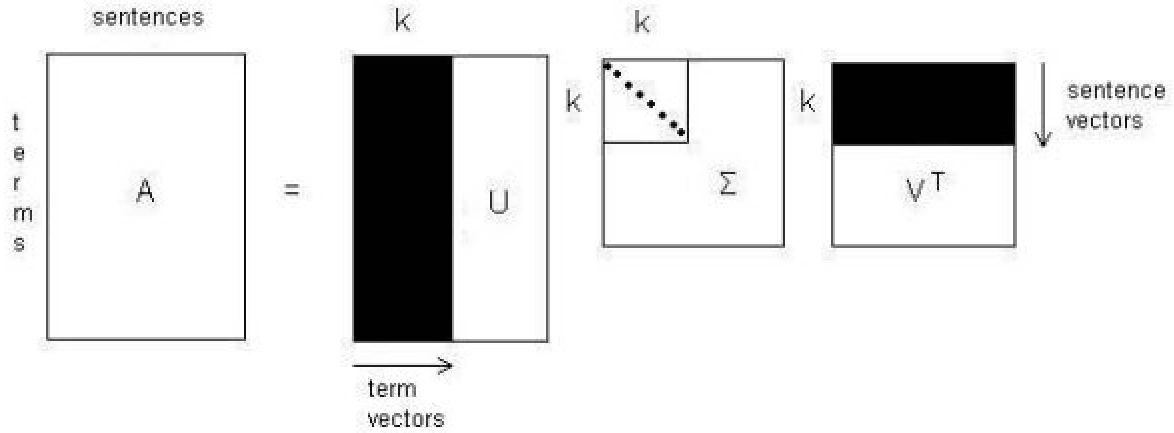
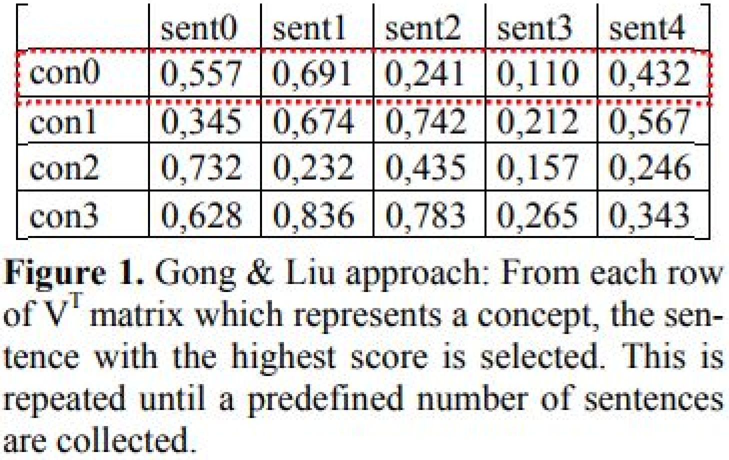


Figure 1: Singular Value Decomposition

LSA and Sentence Selection

- Input Matrix Creation
- Singular Value Decomposition
- Sentence Selection



	sent0	sent1	sent2	sent3	sent4
con0	0,557	0,691	0,241	0,110	0,432
con1	0,345	0,674	0,742	0,212	0,567
con2	0,732	0,232	0,435	0,157	0,246
con3	0,628	0,836	0,783	0,265	0,343

Figure 1. Gong & Liu approach: From each row of V^T matrix which represents a concept, the sentence with the highest score is selected. This is repeated until a predefined number of sentences are collected.

Approach

- Text Pre-processing - Tokenization, Stopwords removal, numeric and punctuation removal, Stemming(if necessary) and PoS Tagging
- Work with either whole sentences or just noun chunks.
- Spell Check
- Sentiment identification
- Sentence Elimination -
  - Retain sentences with only nouns, only adjectives, nouns and adjectives.
  - Retain sentence with either positive orientiation or negative orientation(can be eliminated) or neutral orientation.
  - Use WordNet to determine cohesive relations between terms and remove any redundant sentences.
  - Retain sentences with only specific words.
- Sentence Categorization (optional) - Proposal 2
- Run LSA Summarizer
- Summary correctness (optional)

Proposal 2 : Multilabel classification of user reviews into relevant categories

- Categorize user reviews into 5 important high level categories- “Food”, “Service”, “Ambience”, “Deals/Discounts”, “Worthiness”
- Why is it a multilabel classification?
- Usefulness: This implementation helps-
  - previous summarization task in identifying the top sentences across each category. Later this can be used in presenting the user with relevant tags from all categories against each restaurant
  - a new customer to make a personalized choice, when he/she does not have much time to spend on reading the reviews.
  - Restaurants can be ranked according to these categories.

Approach

1. Binary classifier for each category
2. Multi-class Classifier for each subset of categories (considers correlations between categories, but large category subsets, in our case 2 power 5)
3. Ensemble of subset classifiers- For example, let's say there are 4 categories {Food, Service, Ambience, Deals}. We choose subset size = 2. Hence, we build a total of $^4C_2$ = 6 classifiers for the following combination of categories: {(Food,Service), (Food,Ambience), (Food, Deals), (Service, Ambience), (Service, Deals), (Ambience, Deals)}. For prediction, we consider prediction of all the six classifiers and then take a majority vote.
4. Any Unsupervised learning techniques considering the semantic attributes of text.

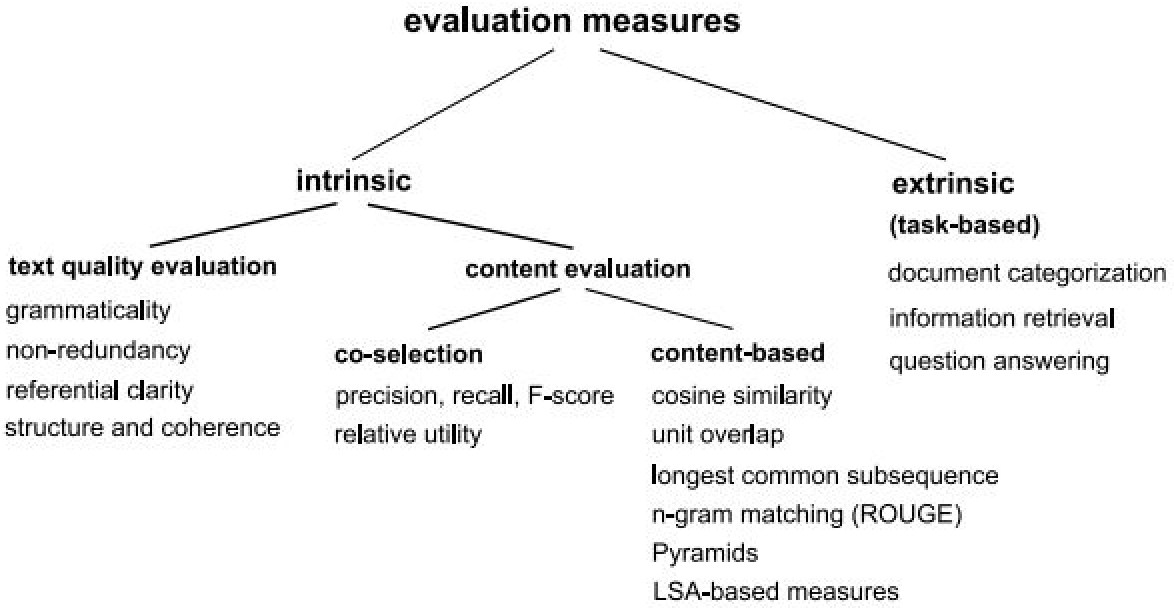
Proposal 3: Evaluation of text summarization results

The evaluation of a summary quality is a very ambitious task.. There are a variety of possible bases for the comparison of summarization systems performance. We can compare a system summary to the source text, to a human-generated summary or to another system summary.

Summarization evaluation methods can be broadly classified into two categories-

1. Extrinsic evaluation- the summary quality is judged on the basis of how helpful summaries are for a given task.
2. Intrinsic evaluation - directly based on analysis of the summary.

Summarization Evaluation Categories



evaluation measures

intrinsic

text quality evaluation

grammaticality

non-redundancy

referential clarity

structure and coherence

content evaluation

co-selection

precision, recall, F-score

relative utility

content-based

cosine similarity

unit overlap

longest common subsequence

n-gram matching (ROUGE)

Pyramids

LSA-based measures

extrinsic
(task-based)

document categorization

information retrieval

question answering

Identifying correct evaluation technique

- **Intrinsic evaluation methods** cannot be applied directly in our case, since we do not have any reference or manually generated summaries on user reviews for restaurants. But this can be achieved if you crawl restaurant summaries from yelp or any other restaurant searching websites.
- **Extrinsic evaluation measures**-

  Document categorization-

  - Here the evaluation seeks to determine whether the generic summary is effective in capturing whatever information in the restaurant all reviews file is needed to correctly categorize the restuarant according to Yelp dataset.

- Information Retrieval approach-
  - Suppose that given query Q and a corpus of documents D, a search engine ranks all documents in D acc. to their relevance to query Q. If instead of corpus D, the corresponding summaries of all documents are substituted the corpus of summaries S is ranked by the same retrieval engine for relevance to the query, the ranking should be almost similar.
  - Few methods used to evaluate similarity of rankings is Kendall’s tau and Spearman’s rank correlation

Resources

For Text Summarization: User reviews of restaurants of 5 cities are crawled from yelp, tripadvisor, zomato already (1m users, 10m reviews for 6 cities around the world using custom crawlers) and combined into text files, one per every restaurant.

For Sentence Categorization: A labelled dataset from Yelp Challenge is available for multilabel classification.

References

[1] Josef Steinberger, Karel Jeˇzek, “Text Summarization using Latent Semantic Analysis”

[2] Josef atSteinberger, Karel Jeˇzek, “Evaluation Measures for text summarization”

[3] http://www.ics.uci.edu/~vpsaini/

[4] Minqing Hu and Bing Liu, “Mining and Summarizing Customer Reviews”

[5] WordNet