Contents

1 Summarizing Conversations

1.1 Introduction

In order to broach the subject of summarizing conversations, a discourse in general Natural Language Processing (NLP) and summarization techniques is in order.

Research Questions

1. How are topics considered off-topic handled?
2. Is the reply structure required in order to appropriately summarize conversations?
    - By what degree does the reply structure improve summaries? i.e. modeling, summary outcomes, etc.
3. What considerations must be taken for two participants vs. N participants?
4. What is the best mechanism for assessing the *correctness* of a summary? Why?

1.1.1 Extraction

Identifies important pieces of text within a corpus (body of text) and builds a summary which contains only those words.

1. Building an Extraction

    Steps

    (a) Construct an IR (Intermediate Representation)

    (b) Score Sentences based on a scoring algorithm

(c) Select a summary based on the scored sentences

(a) IR

i. Topic Representation
Interprets the topics discussed in the corpus. May use a Frequency approach, sentiment analysis, topic word (dictionary), or Bayesian Topic Model approach.

A. Frequency Frequency of words used to determine a *topic*. This can be taken a step further by using the Log-Likelihood Ratio Test.

ii. Indicator Representation
Describes every sentence as a list with important covariates such as word count, length, position in the document, and presence of keywords.

(b) Sentence Score

In a topic IR, this score is an indicator of importance of the sentence. In an Indicator IR, this score is some model based off the covariates. Importance of a sentence can be determined either by **count** or **proportion** of topic words.

(c) Summary Selection

Selects the k most important sentences for the summary. Additional criteria beyond the score may be assessed to determine the sentences chosen for the summary. i.e. Type of document (Newspaper, Blog, Magazine, etc.)

i. Topic Representation

A. Frequency Approach

B. Word Probability
Probability of a word occuring in a document.
$P(w) = \frac{f(w)}{N}$
For each sentence, the average probability of a word is assigned as a *weight*. Then, the best scoring sentence with the highest probability word is chosen to ensure that the sentence is present in the summary. The weight of the chosen word is then updated to ensure that a word in the summary is not chosen over a word that only occurs once [1].
$p_{new}(w_i) = p_{old}(w_i)p_{old}(w_i)$

[1] Unsure of this in particular. Need confirmation

C. Term Frequency Inverse Document Frequency (TFIDF)
A weighting technique which penalizes words that occur most frequently in a document.
$q(w) = f_d(w) \log(\frac{|D|}{f_D(w)})$
$f_d(w)$: Term frequency of a word (w) in a document (d)
$f_D(w)$: Number of documents that contain the word (w)
$|D|$: Number of documents in a collection (D)

- easy and fast to compute.
- Used in many text summarizers

D. Centroid-based Summarization A method of ranking sentences based on TFIDF
Steps

E. Detect Topics and documents that describe the same topic clustered together

- TFIDF vectors are calculated and TFIDF scores below a predefined threshold are removed

F. Clustering Algorithm is run over TFIDF vectors and centroids (median of a cluster) are recomputed after each document is added.

- Centroids may be considered pseudo-documents which contain a higher than the predefined TFIDF threshold.

G. Use Centroids to find sentences related to the topic central to the cluster

- Cluster-based Relative Utility (CBRU) describes how relevant the topic is to the general topic of the cluster.
- Cross Sentence Informational Subsumption (CSIS) measures redundancy between sentences

H. Latent Semantic Analysis
Unsupervised method to selected highly ranked sentences for single and multi-document summaries. Let an n x m matrix exist where n_i is a word in the corpus and m_j is a sentence. Each entry a_{ij} is the TFIDF weight for given word and sentence. Singular Value Decomposition (SVD) is then applied to retrieve three matrices: $A = U\Sigma V^T$ where $D = \Sigma V^T$ describes the relationship between a sentence and a topic.
The assumption is that a topic can be expressed in a single sentence which is not always the case. Additional al-

ternatives have been suggested to overcome this assumption.

I. Bayesian Topic Models
Using probability distributions to model probability of words overcomes two limitations present in other methods:

J. Sentences are assumed to be independent so topics embedded in documents are ignored

K. Sentence scores are heuristics and therefore hard to interpret

The scoring used in Bayesian topic models is typically the Kullbak-Liebler (KL) which measures the difference between two probability distributions P and Q.

ii. Indicator Representation

A. Graph
Represent documents as a graph. Often influenced by PageRank. Sentences are the vertices and edges are similarity (weights). Most common weight is cosine similarity against TFIDF weights for given words.

B. Machine Learning
Approach summarization as a classification problem. Machine Learning techniques include:

- Naive Bayes
- Decision Trees
- Support Vector Machines
- Hidden Markov Models*
- Conditional Random Fields*

*Assume Dependence

Models that assume dependence often outperform those who do not.

2. Abstraction

Interprets and analyzes important pieces of text within a corpus and builds a human readable summary. This is more advanced and computation-intenseive than Extraction.

3. Evaluating Summaries

   Principles in evaluating whether a summary is good or not

   (a) Decide and specify the most important parts of the original text

   (b) Identify important info in the candidate summary since the information can be represented using disparate expressions.

   (c) Readability

   (a) Human Evaluation

   Self explanatory.

   (b) Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

   Determine the quality of a summary by comparing it to human summaries.

   i. ROUGE-n

   gram: a word

   A series of n-grams is created from the reference summary and the candidate summary (usually 2-3 and rarely 4 grams).

   p = number of common n-grams

   q = number of n-grams from reference summary

$$ROUGE - n = \frac{p}{q}$$

   ii. ROUGE-l

   Longest Common Subsequence (LCS) betweeen two sequences of text. The longer the LCS, the more similar they are. Requires ordering to be the same.

   iii. ROUGE-SU

   Also called *skip-bi-gram* and *uni-gram*.

   Allows insertion of words between the first and last words of bi-grams so consecutive words are not needed unlike ROUGE-n and ROUGE-l.

1.2 Conversations

Indicative vs Informative Summarization

Informative: A concise replacement for one or more documents

Indicative: Provides an idea about what is discussed in a document opposed to replacing it. Aims to help the reader assess if the conversation is worth reading.

Current research suggests that Indicative summaries created by Extraction techniques are most appropriate. This is because using Abstraction to generate *headlines* is a hard[2] problem.

The more *focused* a conversation is, the easier it is to process. A focused conversation contains more informative information in the root or lower depth of a tree. Less focused conversations will have larger depths and are more likely have better information lower in the tree[3]. For focused converations, one sentence per message is needed. This can be further optimized by including one from the root and one from the first leaf.

Features to assess

- Depth of discussion Tree
- Branches
- Subject
  - If present, it can be used to score sentences that relate the most. Similar to ROUGE-n

Message Cleansing

- Remove signature blocks and quotes
- Other cleansing needed?

Normalizing sentences gives preference to short sentences; however, the summaries created by the shorter sentences were not necessarily the most accurate or appropriate.

Things to look up

- Non-inflected lexical form from Word-Net?
- LT POS Algorithm
- Visualization Techniques. Graph?

[2] Is hard referring to NP-Hard or Hard as in difficult?

[3] Need more proof here.

2 References

- Brief Introduction to NLP
- Overview of Text Summarization Techniques
  - See Section 5 for further references to review for conversation summaries
  - **Nathan**: See section 7
  - 45 - summary for 2 levels of discussion
  - 56 - ML with features
  - 47 - summarize a full mailbox rather than a thread with clustering and extracting summaries for each cluster
- Facilitating Email Thread Access by Extractive Summary Generation
  - Newman 02a-b - Characterize all threads. Same as 47?
  - Kan 01 - Deeper dive on Indicative vs Informative Summaries