Contents

1 Summarization

1.1 Extraction

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

1.1.1 Building an Extraction

Steps

1. Construct an IR (Intermediate Representation)
2. Score Sentences based on a scoring algorithm
3. Select a summary based on the scored sentences

1. IR

   (a) Topic Representation

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

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

(b) Indicator Representation

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

2. 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.

3. 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.)

(a) Topic Representation

i. Frequency Approach

A. 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].

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

B. Term Frequency Inverse Document Frequency (TFIDF)

A weighting technique which penalizes words that occur most frequently in a document.

$q(w) = f_d(w) \log(\frac{|D|}{f_D(w)})$

[1] Unsure of this in particular. Need confirmation

$f_d(w)$: 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

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

D. Detect Topics and documents that describe the same topic clustered together
   - TFIDF vectors are calculated and TFIDF scores below a predefined threshold are removed

E. 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.

F. 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

ii. 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 alternatives have been suggested to overcome this assumption.

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

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

B. 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.

(b) Indicator Representation

i. 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.

ii. 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.

1.2 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.

1.3 Evaluating Summaries

Principles in evaluating whether a summary is good or not

1. Decide and specify the most important parts of the original text

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

3. Readability

1.3.1 Human Evaluation

Self explanatory.

1.3.2 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

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

1. 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

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

2. 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.

3. 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.

2 References

2.1 Brief Introduction to NLP

2.2 Overview of Text Summarization Techniques

2.2.1 See Section 5 for further references to review for conversation summaries

2.2.2 Nathan: See section 7