Contents

- Goal of summarization is generally considered to be: create summary which is most similar to one generated by a human
- Abstractive vs Extractive Summarization
  - Extractive simply extracts summary sentences verbatim from corpus
  - Abstractive generates new text from corpus
  - Almost all summarization methods are extractive
  - Humans create abstractive summaries, but these are an order of magnitude more complex.
- Historically, the methods used in text summarization are very closely related to those used in IR.
- Some of the concepts we've looked at for IR that have been used in automatic summarization include:
  - Frequency driven approaches using TFIDF
  - Clustering ala ROCCHIO
  - Naive Bayes
  - PageRank inspired

1 Frequency Driven Approach

- The weight of each word w in document d is computed by:

$$q(w) = f_d(w) * log \frac{|D|}{f_D(w)} \tag{1}$$

where $f_d(w)$ is the frequency of word w in document d $f_D(w)$ is the number of documents that contain w, and |D| is the number of documents

- A variety of techniques make use of this weighting scheme
- One fairly ubiquitous summarization algorithm is SumBasic.
  - Each sentence is assigned a weight from the average weights of the words in that sentence.
  - The highest weight sentence which contains the highest weight word (topic word) is chosen

2 Clustering

- Used to derive topics and topic importance
- Sentences are clustered from TFIDF vector representation, often low-weight sentences are filtered out.
- Clusters with many sentences are considered more important topics
- From here, each cluster can be treated as a document. Summaries can then be generated by traditional summary techniques.

3 PageRank inspired graphical algorithms

- Represent sentences/documents as nodes
- Create edges based on sentences which pass a chosen similarity threshold
  - Often cosine similarity from TFIDF vector representation
- Nodes with many edges are considered more important, and more likely to be chosen for extractive summaries
- Additionally, the structure of the graph could be used to determine topics (by examining sub-graphs)

4 Naive Bayes

- Naive Bayes Classifier can be used in a machine learning approach.

P(summary sentence | words in sentence) can be approximated from training data.

5 Other Approaches

- Bayesian Topic Model using Kullbak-Liebler Divergence
- Machine Learning approaches
  - Machine Learning solutions show widespread success in a variety of areas
  - Superficially, we have access to large amounts of data, the main prerequisite for most machine learning approaches
  - Unfortunately, this data does not include labeled summaries (in general)
- Ontologies
  - Manually created for specific domains e.g. UMLS for medical
  - Automatically generated e.g. YAGO generated from wikipedia articles

6 Proposed Algorithm

- Proposed Algorithm for Generating a concise summary from a large, general corpus:
  - Assign weight to every document using graph-based approach
  - Vector-space model, use cosign similarity with query to select subset of documents
  - Cluster documents using ROCCHIO to derive subtopics
  - Select top n documents from each cluster based on graph-based weighting
  - Compute Probability Distribution P over words w for each cluster.
  - For each cluster extract sentences using Kullbak-Liebler Divergence.
  - Concatenate topic summaries

7 Evaluation

- Historically, a large amount of summarization research has occurred at summarization-specific conferences where human judges perform evaluation.
- Most common automatic evaluation is ROUGE (Recall Oriented Understudy for Gisting Evaluation), but these methods still requires human generated summaries for comparison.
- There are many variation of ROUGE, but some common ones include:
- ROUGE-n : based on comparison of n-grams. Let p be the number of common n-grams between candidate and reference summary, and q b the number of n-grams from the reference summary only, ROUGE-N = p/q
- ROUGE-L : based on longest common subsequence between candidate and reference summary
- Other studies perform ad-hoc evaluation using metrics from IR.

8 Looking forward

- Newer methods for text summarization prefer methods from natural language processing over those from IR
- These methods tends to be more complex and more computationally expensive
- Examples include more sophisticated encoding of documents/sentences/words using neural networks
- Additionally there have been gains from using semantic analysis thanks to resources such as WordNet