Abstract

Text summarization is a data reduction process. The use of text summarization enables users to reduce the amount of text that must be read while still assimilating the core information. The data reduction offered by text summarization is particularly useful in the biomedical domain, where physicians must continuously find clinical trial study information to incorporate into their patient treatment efforts. Such efforts are often hampered by the high-volume of publications.

Our contribution is two-fold: 1) to propose the frequency of domain concepts as a method to identify important sentences within a full-text; and 2) propose a new model and algorithm for identifying important sentences based on term or concept frequency distribution. It is shown that the use of concepts performs closely with the use of terms for sentence selection. Our proposed frequency distribution model algorithm outperforms a state-of-the-art approach.

Motivation, Hypothesis and Method

Motivation: Generate short summaries of biomedical texts (randomized controlled trials in oncology) to allow physicians and researchers to assimilate more information in less time.

Approach: Identify and extract sentences from the source text to form a summary. The problem is how to identify important sentences within the full source text to extract while simultaneously reducing information redundancy.

Hypothesis: The distribution of concepts within a summary should approximate the same distribution of concepts in the full source.

1. No ideal summary exists
2. Abstract is one summary from one viewpoint
3. Does not consider user’s information need
4. Abstract may be missing content from full-text
5. Generation of customized content
6. Question-answering systems
7. Evaluate sentence selection methods

Method: Build a summary incrementally by adding a sentence to the summary and then comparing the summary concept distribution to the source text concept distribution. To compare similarity between summary and source concept distribution, the distributions for each are modeled as vectors, and then compared using the methods below:

Evaluation

Basic idea
- Generate ideal (’model’) summaries from domain experts
- Generate system summaries and compare to model summaries

Resources
- A corpus of 24 randomly selected full biomedical texts was used
- Three domain experts generated extractive model summaries for each paper (20% compression)
- Six system summarizers generated extractive summaries (20% compression), used for comparison

ROUGE n-gram overlap tool
- ROUGE: Recall Oriented Understudy for Gisting
- ROUGE: Generates summaries from model summaries
- Results based on n-gram overlap
- ROUGE-2: bigram co-occurrence
- ROUGE-SU4: skip bigram with no more than 4 intervening words
- Same metrics as used in 2005 Document Understanding Conference

Results & Discussion

Observations:
- Random summarizer performs relatively well
- Ignoring FreqDist, Baseline Lead is worst performing
- Unlike news genre, where Lead is competitive
- Use of Dice for FreqDist works best
- Use of terms and concepts perform closely
- But concepts can allow for easier tailoring of a summary

Lawrence H. Reeve1, Hyoil Han1, Saya V. Nagori1, Jonathan C. Yang2, Tamara A. Schwimmer3, Ari D. Brooks2
1 Drexel University, College of Information Science and Technology
2 Drexel University, College of Medicine

References
- Motivation, Hypothesis and Method
- Evaluation
- Results & Discussion

Concept Frequency Distribution in Biomedical Text Summarization

Lawrence H. Reeve1, Hyoil Han1, Saya V. Nagori1, Jonathan C. Yang2, Tamara A. Schwimmer3, Ari D. Brooks2
1 Drexel University, College of Information Science and Technology
2 Drexel University, College of Medicine

1. Domain-Specific Resource

The domain-specific resource provides a thesaurus for a particular domain. The thesaurus has a list of concepts and for each concept, a list of synonymous phrases. The synonymous phrases are used to find concepts within the source text.

In this example, the concept ‘Multiple Myeloma’ has five synonymous phrases which may be found in biomedical text.

The research uses the Unified Medical Language System (UMLS) Metathesaurus, which contains over 100 different vocabulary sources.

2. Source Text

The source text used is a randomized controlled trial journal paper. Although the biomedical domain is used, any domain providing domain-specific resources can be used.

3. Extractive Text Summarization

Domain concept frequency distribution follows a Zipf distribution. Biomedical text concept distribution across corpus of 24 papers, a shows a distribution of 4267 concepts across medical concepts within the papers. The FreqDist algorithm uses the Zipf distribution observation to match summary and full-text:

- Build frequency model of full-text
- Iteratively select sentences from full-text:
  - Take a sentence, add it to current summary. This is a candidate summary
  - Compare candidate summary distribution model for similarity to full-text distribution model
  - Select sentence which, when added to summary, best models the full-text

4. Generated Text Summary

To determine the summary size, the user specifies a percentage of the original source text to return (e.g., 20%) or a number of sentences. The summary can be tailored by weighting concepts.

Figure 1: Text Summarization from Concepts Using Frequency Distribution

Figure 2: Concept Similarity Functions evaluated

Figure 3: ROUGE Scores