CIS 430/530 Final Project: Summarization

Jason S. Mow
jmow@seas.upenn.edu

Nate Close
closen@seas.upenn.edu

1 Basic Systems

Below we illustrate our approaches and choice of parameters for the three basic summarization systems.

1.1 Centroid-Based System

The Centrality Summarizer has the following parameters to configure it's functionality:

- Vector Feature Weight Representation

- Similarity Comparison Approach

- Sentence Length Limits (short and long)

- Redundancy Removal Approach

We chose a binary representation for our sentence vector feature weight. We did this because it was the simplest to compute and yielded strong results in our preliminary testing. Our similarity approach was to use cosine similarity on the sentence vectors.

Our sentence length limit was between 15 and 50 words, tokenized by NLTK. We mitigated redundancy by rejecting any sentences with a cosine similarity greater than 0.75 with any sentence already in the summary. Again, these thresholds were chosen as they yielded the most sensible and well-scoring results from our trials.

1.2 Topic-word Based System

The Topic-Word Summarizer has the following parameters to configure it's functionality:

- Sentence Score Normalization

- Topic Word Cutoff

- Sentence Length Limits (short and long)

- Redundancy Removal Approach

For sentence vector feature weight, we chose the third representation which calculates weight as (# of topic words / # of nonstopwords). This choice seemed most logical to us as it doesn't dilute the score with stopwords, and also normalizes for sentence length.

Topic Word Cutoff was set to 0.1. This was the default setting and was not adjusted / tested extensively due to time constraints with regenerating topic word files. We deemed this to be an optimal setting after testing out the results on several different cutoff thresholds.

Our sentence length limit was between 15 and 50 words, tokenized by NLTK. We mitigated redundancy by rejecting any sentences with a cosine similarity greater than 0.75 with any sentence already in the summary. Again, these thresholds were chosen as they yielded the most sensible and well-scoring results from our trials.

1.3 LexPageRank System

The LexPageRank Summarizer has the following parameters to configure it's functionality:

- Edge Similarity Threshold

- LexRank End Criteria - ¡ 0.001 change

- Sentence Length Limits (short and long)

- Redundancy Removal Approach

For the LexRank summarizer, we chose to use TF-IDF representation over binary representation. This produced more accurate vectors and better results from ROUGE in the summarization.

For edge similarity threshold, we chose the value of 0.2. This was suggested in the reference text discussing Lex Page Rank, and we found it to be fairly successful. For this value, too, we had limited ability to vary and continue to experiment as the process of generating summaries was extremely time-consuming.

Our LexRank End Criteria was set such that the iteration would end if all values changed less than 0.001 between iterations. This was a good medium between performance and getting reasonable results. Also, the results did not change much as the threshold was decreased further.

Our sentence length limit was between 15 and 50 words, tokenized by NLTK. We mitigated redundancy by rejecting any sentences with a cosine similarity greater than 0.75 with any sentence already in the summary. Again, these thresholds were chosen as they yielded the most sensible and well-scoring results from our trials.

1.4 Performance

2 Custom Summarization System

2.1 System Design

This part shows general idea of your system. You may use flowchart, graphics or pseudo-code to describe your algorithm.

2.2 Resources & Tools Used

What resources or tools you have used and how they are included in your implementations.

Example: Wordnet, Stanford-Parser, MPQA.

I use Stanford-Parser in order to help

2.3 Performance

3 Discussion and Analysis