Ling 573 Summarization Presentation

Thomas Marsh
University of Washington
ling 573, Spring 2015
sugarork@uw.edu

Michael Roylance
University of Washington
ling 573, Spring 2015
roylance@uw.edu

Brandon Gahler
University of Washington
ling 573, Spring 2015
bjg6@uw.edu

Abstract

TODO: fill out

1 Introduction

TODO: fill out

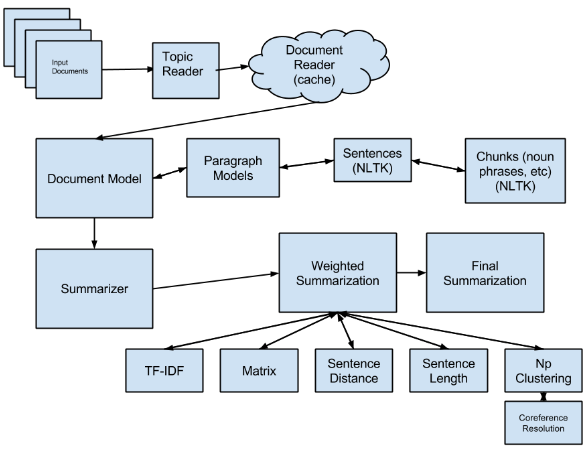
2 System Overview Diagram

TODO: fill out

3 Approach

3.1 System Architecture

Figure 1: System Architecture



3.2 Content Selection

We implemented a variety of methods initially to help with content selection. Each of these five separately implemented content selection systems produces a score for each sentence from 0.0 to 1.0. Empirically determined system weights are used in summing these scores to produce aggregate scores. The top scoring sentences are used to create the summary itself.

3.2.1 Trivial Systems

Two of our content selection systems are trivial systems. One system scores sentences based on their length relative to the longest sentence in the document cluster, favoring longer sentences. The second scores sentences based on their position in their document, giving the first sentence in each document a score of 1.0 and the last sentence in each document a score of 0.0.

3.2.2 TF-IDF Scoring

For our first non-trivial content selection system, we calculated the average tf-idf score of every non-stop word in the sentence. TF-IDF scores where calculated using the Reuters-21578, Distribution 1.0 Corpus of news articles, as incorporated in the NLTK, as our background corpus. Sentence scores are scaled such that the sentence with the highest average TF-IDF was given a score of 1.0.

3.2.3 Simple Graph Based Scoring

Next, we implemented a simple graph based metric, in which a dense undirected graph of sentences is constructed with edge weights set to the cosine similarity of the sentences. The most connected sentence is iteratively selected, and the weights of the edges of the previously selected sentences are set to be negative to discourage redundancy. The first pulled sentence is given a score of 1.0, with scores incrementally decreasing to 0.0 for the sentence pulled last.

3.2.4 Noun Phrase Clustering

Finally, we also implemented a noun phrase clustering algorithm where we used a custom coreference resolution system to resolve pronouns to their most likely antecedent, and then compared each sentence's noun phrases to each other sentence. NP-clustering selection was ordered by the number of matches each sentence had to every other sentence.

3.3 Information Ordering

In this deliverable we did not attempt any information ordering, leaving the summary sentences in the order of importance as determined by the content selection system.

3.4 Content Realization

Our realization was simple, the highest ranking sentences were realized into the summary with only extra white space an newlines removed. We iterated through the aggregate scored sentences in order, adding each sentence if their was room for it left within the 100 word length limit.

3.5 Best Technique

The system we turned in for D2 uses only tf-idf to score sentences, as we though at the time of submission that this was the best system. In the time between system submission and submission of this report we uncovered a bug in our empirical weight generation system. The best technique we found after fixing this problem had the tf-idf system and simple graph system weighted equally.

4 Results

Our results for the submitted tf-idf only solution are shown in Table 1.

Table 1: tf-idf

Rouge Technique	Recall	Precision	F-Score
ROUGE1	0.55024	0.52418	0.53571
ROUGE2	0.44809	0.42604	0.43580
ROUGE3	0.38723	0.36788	0.37643
ROUGE4	0.33438	0.31742	0.32490

Our best scoring system was tf-idf enhanced with our simple graph system (Table 2), followed by the simple graph by itself (Table 3). The simple graph system notably is more precision oriented than the tf-idf system, with the aggregate producing higher F-scores than tf-idf despite lower recall. Lastly, though it was not ultimately successful, we have included the results for the NP clustering technique (Table 4) as well.

Table 2: tf-idf + Simple Graph

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.54107	0.57388	0.55580
ROUGE-2	0.42822	0.45443	0.43997
ROUGE-3	0.36791	0.39088	0.37819
ROUGE-4	0.31867	0.33882	0.32767

Table 3: Simple Graph

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.48228	0.56860	0.52048
ROUGE-2	0.36821	0.43541	0.39787
ROUGE-3	0.31484	0.37348	0.34065
ROUGE-4	0.27465	0.32683	0.29757

Table 4: NP Clustering

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.45691	0.53378	0.49056
ROUGE-2	0.33306	0.39053	0.35813
ROUGE-3	0.28221	0.33196	0.30386
ROUGE-4	0.24758	0.29237	0.26700

4.1 Error Analysis

There are many things to still tweak with both np-clustering and the matrix comparison. For example, we could do better normalization with other co-referents rather than just pronouns. We will be experimenting with these techniques further.

5 Discussion

TODO: fill out

6 Conclusion

Conclusions have been made as can be seen from the following Nenkova, Radev, and Jones (Nenkova et al., 2007) (Jones, 2007) (Radev et al., 2001). We used co-referenced based off ideas from (Cardie and Wagstaff, 1999).

References

C. Cardie and K. Wagstaff. 1999. Noun phrase coreference as clustering. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1001:82–89.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing Management*, 43(6):1449 – 1481. Text Summarization.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4.

Dragomir R Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multi-document summarization using mead. *Ann Arbor*, 1001:48109.