Ling 573 Summarization Presentation

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

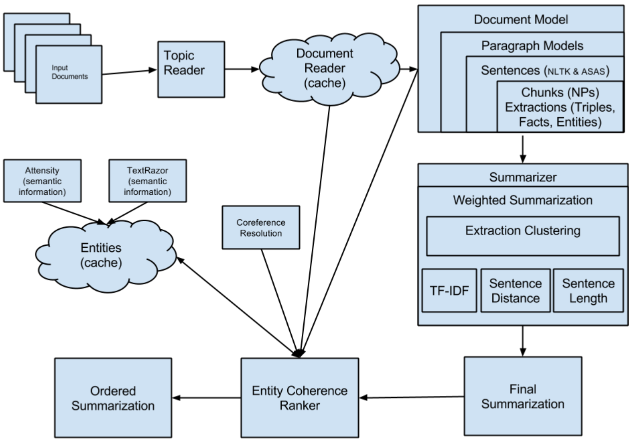
Michael Roylance
University of Washington
ling 573, Spring 2015
roylance@uw.edu

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

Abstract

This research looks into several different techniques for automatically summarizing a group of documents. It explores simple sentence similarity techniques such as TF-IDF, sentence distance and sentence length, and then delves into slightly more complicated approaches such as graph similarity, noun phrase clustering, and an extracted ”subject predicate object” case frames and entity (from named entity recognition) clustering (Riloff and Schmelzenbach, 1998). Finally it attempts to improve the score by reordering the sentences with entity sequence learning (Barzilay and Lapata, 2008).

Figure 1: System Architecture



1 Introduction

Text summarization has the potential to be a powerful technique for displaying the most relevant information in just a few characters. As a form of a question/answering system, it can effectively communicate complex information that would otherwise take a person much longer to read. In this paper, we explore several techniques.

2 Approach

2.1 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.

2.1.1 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.

2.1.2 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.

2.1.3 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.

2.1.4 Noun Phrase Clustering

We also implemented a noun phrase clustering algorithm where we used a custom co-reference 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.

2.1.5 Extraction Clustering

Finally, we decided to look at external sources for information extraction use in sentence similarity and clustering.

The first extraction we used was entities, which contain semantic information for identified noun phrases (person, location, company, etc). For sentence similarity identification, we compared not only the noun phrase spelling, but also the semantic tag associated with it

The second extraction we used was an extraction called ”triples”. The parser identifies several subject, predicate and object tuples in each sentence. For our similarity comparison, we look at just the subject and object.

The final extraction we used from the parser was called ”facts”. Facts have two properties associated with them, ”element” and ”mode”. These two properties combine with other extractions (not used for this project, currently) similar to triples for generating case frames (Riloff and Schmelzenbach, 1998)

2.2 Information Ordering

In this deliverable, we implemented a version of information ordering based on (Barzilay and Lapata, 2008).

Specifically, we used named entity recognizers to extract entities to use in transition grids. Although we do have a coreference resolution system in place, we did not have time to hook it up properly to this information ordering system in time for deliverable three.

The entity coherence system reads in source documents and extracts entities from two commercial products, Text Razor and Attensity. It assigns each entity a type of Subject, Object or Oblique. Both parsers assign freebase ids to each entity, assuring global uniqueness. We used the Reuters corpus for training on correct ordering, extracting the feature vectors from the current ordering then an induced randomly ordering.

We then do ranking training on the feature vectors with SVMLight and reorder our current sentences based on the maximum score from the training model.

2.3 Content Realization

Our realization was simple, the highest ranking sentences were added into the summary with only extra white space and newlines removed. We iterated through the aggregate scored sentences in order, adding each sentence if there was room left within the one hundred word length limit.

2.4 Best Technique

The best system turned out to be the noun phrase clustering one, followed closely by an ”equal weight” between all the systems.

3 Results

Our best results are shown below in Table 1. It turned out to be a close race, with our NP-Clustering-only solution (Table 1) winning over an equal weight system (Table 2) by a slim margin. We also ran the tf-idf by itself (Table 4), as well as Simple Graph only (Table 3). We found that our techniques generally worked the best when working together, but the NP Clustering by itself seemed to be the exception.

Table 1: NP Clustering

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.23391	0.28553	0.25522
ROUGE-2	0.05736	0.07053	0.06272
ROUGE-3	0.01612	0.01969	0.01758
ROUGE-4	0.00533	0.00657	0.00584

Table 2: All Techniques

Rouge Technique	Recall	Precision	F-Score
ROUGE1	0.23336	0.28628	0.25516
ROUGE2	0.05708	0.07044	0.06251
ROUGE3	0.01612	0.01969	0.01758
ROUGE4	0.00533	0.00657	0.00584

Table 3: Simple Graph

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.19379	0.25550	0.21845
ROUGE-2	0.04473	0.05859	0.05033
ROUGE-3	0.01170	0.01505	0.01305
ROUGE-4	0.00362	0.00453	0.00400

Table 4: tf-idf

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.15341	0.20846	0.17522
ROUGE-2	0.03014	0.04037	0.03426
ROUGE-3	0.00746	0.01038	0.00863
ROUGE-4	0.00242	0.00329	0.00278

Table 5: Extraction Clustering

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.23582	0.24725	0.24071
ROUGE-2	0.06255	0.06447	0.06337
ROUGE-3	0.01995	0.02048	0.02018
ROUGE-4	0.00659	0.00672	0.00664

Table 5: Extraction Clustering with Entity Reordering

Rouge Technique	Recall	Precision	F-Score
ROUGE-1	0.22494	0.27984	0.24734
ROUGE-2	0.05357	0.06688	0.05893
ROUGE-3	0.01532	0.01904	0.01683
ROUGE-4	0.00491	0.00614	0.00541

3.1 Error Analysis

While our score did improve, there are several things that we can do for improvement.

1. Our selection algorithm is grabbing sentences with many keywords, but they don't seem to be communicating the overall summary of the document. Specifically, some of our summaries make little sense. Our ROUGE2 score did increase, however.

2. Our sentence ordering algorithm will be improved with different corpora

3. It seems that we should look into applying a heuristic based approach for some of our content selection as well. This could be useful in removing unnecessary text from sentences like newspaper annotations

4 Discussion

We are encouraged with our new clustering results, however, we will continue to look at improving our ordering technique for better coherence.

5 Conclusion

We have seen improved ROUGE-2 scores with extraction based clustering (Riloff and Schmelzenbach, 1998). We are still working to improve the cohesion scores. 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

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

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.

Ellen Riloff and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition.