LING 573 - MultiDocument Summarization System

Audrey Holmes
University of Washington
auholmes@uw.edu

Eslam Elsawy
University of Washington
eslam@uw.edu

Masha Ivenskaya
University of Washington
marliven@uw.edu

Abstract

1 Introduction

2 System overview

Figure 1 shows the architecture of the baseline summarization system. First, sentences from documents belonging to the same topic are extracted. Then, adjacency matrix is built with pairwise tf-idf cosine similarities. After that, PageRank algorithm utilizes the similarity matrix to sort the sentences based on their importance. Finally, sentences and importance scores are fed into the reranker module which picks the sentences to consider in the final summary.

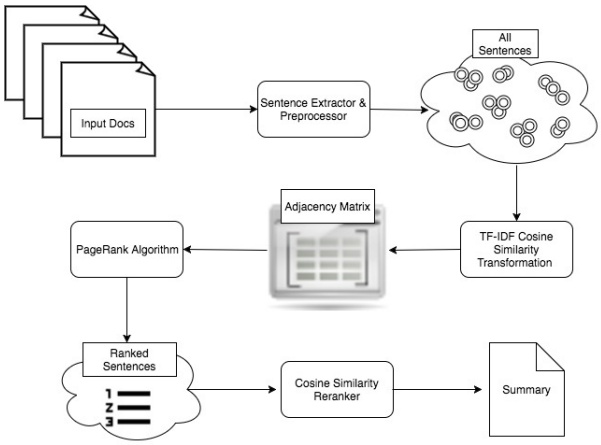

Figure 1: System architecture

3 Approach

3.1 Data

We use the TAC-2010 dataset that consists of the following:

- A set of n topics $T = \{T_1, T_2, ..., T_n\}$.
- For each topic T_i, a set of m documents $D_i = \{d_{i,1}, d_{i,2}, ..., d_{i,m}\}$.
- For each topic T_i a vocabulary V_i.

3.2 Content Selection

Sentence Extraction and Preprocessing. For each document, we extract the contents of each document from the XML file using Python's BeautifulSoup module. We then convert the text to lowercase and perform sentence tokenization.

Sentence Similarity. For each document $D_{i,j}$, we convert the sentences into vectors s of length $|V_i|$, where s_v is the number of times word v appears in the sentence. From these vectors, we create an N x $|V_i|$ matrix M_i, where N is the number of sentences within a topic. We then perform a tf-idf transformation on M_i. Finally, we transform M_i into a symmetric N x N matrix of pairwise cosine similarities.

PageRank. We binarize the entries of M_i such that values less than 0.15 are set to 0; all other values are set to 1. We then implement the Google PageRank (Page et al., 1999) (Erkan and Radev, 2004) to estimate sentence importance.

Other Similarity Measures. In addition to tf-idf cosine similarity, we experimented with using Doc2Vec to measure sentence similarity. We trained a doc2vec model on the NLTK Reuters corpus and then used this model to calculate similarity scores between sentences. These scores were then binarized and ran through the PageRank algorithm. We experimented with different binarization thresholds, as well as with removing stopwords from the sentences before measuring similarity. However the Rouge scores obtained through this approach were lower than the results obtained using tf-idf cosine similarity.

Table 1: ROUGE-1 evaluation results for different baseline system configurations

Similarity Measure	Information Ordering	Precision	Recall	F-score
TF-IDF Cosine Similarity	Dummy	.29	.24	.26
Doc2Vec	Dummy	.18	.23	.20
TF-IDF Cosine Similarity	Cosine Reranker Thr=0.2	**.3**	**.26**	**.28**
Doc2Vec	Cosine Reranker Thr=0.2	.20	.22	.21

Table 2: ROUGE-L evaluation scores for baseline

ROUGE-L	Precision	Recall	F-Score
ROUGE-1	0.3	0.26	0.28
ROUGE-2	0.08	0.07	0.08
ROUGE-3	0.03	0.02	0.03
ROUGE-4	0.01	0.01	0.01

Our hypothesis is that the contextual similarity captured by doc2vec is not the right match for the document summarization task - i.e. if we are measuring how connected a sentence is to other sentences in the topic, we are interested in actual overlap of entities and concepts, not broad semantic relatedness (i.e. a sentence about dogs being related to a sentence about cats).

4 Discussion

4.1 Information Ordering

For our baseline approach, we ordered the sentences by descending PageRank score and cut off each summary at the maximum word limit. We also implemented a cosine similarity reranker (Radev et al.,), skipping sentences that are too similar to sentences already in the summary.

Figure 2 shows that the best achieved results happen at cosine similarity threshold = 0.2

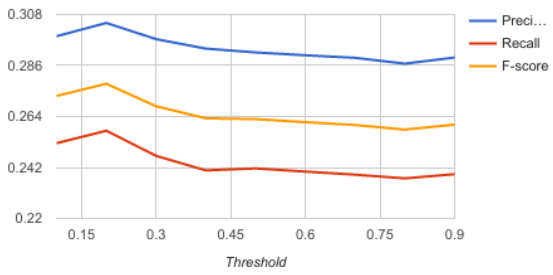

Figure 2: Precision, Recall and F-measure at different reranker cosine similarity thresholds

4.2 Content Realization

We truncate the summaries so that they do not exceed 100 words.

5 Results

Table 1 summarizes our results using different similarity measures and reordering algorithms. The best achieved results occurred when using TF-IDF cosine similarity and cosine sentences reranker with threshold = 0.2.

Table 2 shows best achieved ROUGE-L (Lin,) evaluation scores for our baseline system.

Next steps will be to improve information ordering and content realization. In particular, we will work on removing redundant sentences from the final summaries. Since we got most of our improvements from improving pre-processing and parameter tuning, we will do more of that in the PageRank portion of the code. We noticed that the gold standard summaries seemed to have shorter sentences than our summaries, so we will look deeper into this to see if there is something we can improve.

6 Conclusion

References

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Dragomir R Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, et al. Mead-a platform for multidocument multilingual text summarization.