Veronica Alejandro

Ethan Glass

LIN 353C: Final Report

Introduction

We implemented and evaluated three different methods of text summarization. We believe that summarization is important because of the rise of information and knowledge. Nowadays, there is constant information being put out into the Internet and it can be overwhelming to have to sift through so much on just one topic. For example, we often get millions of search results with each result having a ton of information on a certain topic. Additionally, websites like Wikipedia often have too much information that most people do not need. Thus, automatic text summarization provides users with the most relevant information without having to rely on manual text summarization, which can take a long time for just one article. With our limited capability, we want to find the most effective algorithm for automatic text summarization.

Methods

In this project, we will focus on extractive text summarization, meaning that we will use sentences already in the article for our summaries. The task is then identifying which collection of sentences best summarize the article. We processed our data by lowercasing everything, sentence and word tokenizing, and removing stop words. We created a frequency dictionary for each word in each sentence.

First, we start with a simple summarization method by taking the first sentence of each article. This served as a good comparison for the success of our other algorithms. This was very simple to implement and only required that we pull the first sentence token from each article.

Then, we created TF-IDF values for each sentence. Using the frequency dictionary, we calculated the TF-IDF of each word in a sentence, with respect to the other sentences in the article. This is to say that the IDF of the word was scored according to the number of sentences in the article it appeared in, not according to the number of articles across our data set that it appeared in. We totaled the TF-IDF word values and assigned that score to our sentences. The three sentences with the highest values were chosen as our summaries.

Then we used vector embeddings with the TextRank algorithm to pull our summary sentences. We used GloVe embeddings that had been trained on the GigaWord-5 data set. This dataset consists of text pulled from news sources and Wikipedia. GloVe provided vectors representing the co-occurrences of a given word with other words. We assigned these vectors to each word in our sentences, then averaged the vectors to create one sentence vector. We calculated the cosine similarities of the sentences and created a similarity matrix. We used networkx, a Python package that applied the TextRank algorithm to create a graph which let us determine the most important sentences. The TextRank algorithm gives higher scores to sentences that are more similar to other sentences. We chose the three highest-ranked sentences for our summary.

Data

We used the CNN/Daily Mail dataset which contains 300,000 articles from both CNN and the Daily Mail and short summaries manually written by the article's author. These articles were collected between 2007 and 2015. Articles were scraped from internet archives and were less than 2,000 tokens long. We used 50 article-summary pairs in our development.

Evaluation

We used the ROUGE evaluation metric to assess our summarization algorithms. ROUGE compares unigram similarity, bigram similarity, and the longest common subsequence for any two pieces of text and returns precision, recall and F1 measures. Given that our data already came with human-generated summaries, we had accessible comparison data. Each machine-generated summary was compared against the relevant human-generated summary. We generated scores for the fifty comparisons and averaged them.

Results

Unigram Comparison

Method	Precision	Recall	F1
First-Sentence	0.340	0.205	0.248
TextRank with Word Embeddings	0.243	0.312	0.262
TF-IDF	0.129	0.202	0.151

Bigram Comparison

Method	Precision	Recall	F1
First-Sentence	0.124	0.075	0.091
TextRank with Word Embeddings	0.079	0.099	0.084
TF-IDF	0.016	0.051	0.024

Longest Common Subsequence

Method	Precision	Recall	F1
First-Sentence	0.309	0.186	0.225

TextRank with Word Embeddings	0.216	0.276	0.232
TF-IDF	0.099	0.227	0.135

Discussion

Precision refers to the ratio of the number of overlapping words with the total number words in the computed summary. So, precision tells us how much of the generated summary was relevant. Recall refers to the ratio of the number of overlapping words with the total number of words in the reference summary. So, recall tells us how much of the reference summary that the generated summary captures. And F1 score is of course a weighted average of these two metrics. Summaries can be found in Appendix A.

The first sentence method scored the highest on precision in each measure. Also notable, the first sentence method provided the shortest summary, by a lot. This explains why the precision of the longer summaries are much lower. Word embeddings scored highest on recall for each metric. It is unclear why, but we can interpret this as meaning that our word embedding summary captured the human-made summary better than any other algorithm. Word embeddings had the highest F1 score in Unigrams and Longest Common Subsequence, and first sentence had the highest in Bigrams.

TF-IDF scored very poorly in every measure. Given that the sentence score was a total of each word's TF-IDF score, longer sentences were favored. This brought down the precision considerably. Although it was still the lowest, its recall scores were much closer to the other methods. Given that so much text was used for the summary, we expected recall to be higher. Given that TF-IDF was calculated with respect to the sentences, IDF played a much greater role in the score. Sentences simply are not usually long enough to have multiple occurrences of non-stopwords. This led to sentences with less frequent words to be captured. Appendix A4 provides a great example in which the last two sentences are not relevant for a summary but were captured nonetheless.

We must consider the limitations of the comparison summary. There are many possible ways to summarize an article. Additionally, using extractive methods might lead to more awkward summaries since we can only use what is in the article. Abstractive methods, in which the model creates sentences without relying on the article sentences, for automatic summarization are worth looking into and might be more effective. Any conclusions about the effectiveness of these algorithms should be cautiously drawn, given that all we know is how similar our generated summaries are to one particular human generated summary.

Labor Division

Veronica did the first sentence and word embedding algorithms and the intermediate report. Ethan did the TF-IDF algorithm and the final report. We feel that labor was fairly divided.

Appendix

A. Summaries

1. *Human Generated Summary*

Experts question if packed out planes are putting passengers at risk. U.S consumer advisory group says minimum space must be stipulated. Safety tests conducted on planes with more leg room than airlines offer.

2. *First Sentence Summary*

Ever noticed how plane seats appear to be getting smaller and smaller?

3. *Word Embeddings Summary*

'In a world where animals have more rights to space and food than humans,' said Charlie Leocha, consumer representative on the committee. But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking? Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased.

4. TF-IDF Summary

This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.