Automatic text summarization using graphs

Branislav Anđelić | Fakultet tehničkih nauka

Motivation

In the current society, there is an abundance of information that we do not need, or rather we need the core information from these sources. For human beings, manually summarizing texts can be a tedious and strenuous task. Using text summarization, we solve the problem of presenting information, from large documents of text, in a compact form.

Problem

- Given a document of text, the goal is to identify and extract a number of most important sentences, which will form the summary.
- The problem of identifying the sentences will be approached with two algorithms: *TextRank* and a noun based algorithm.
- Choosing the most important sentences will be done in two ways: trivial (*n* best ranked) and using *uniform-cost search (UCS)*.

Preprocessing

The common steps in preprocessing the text for both algorithms are:

- Loading the text from a file
- Sentence separation
- Word tokenization of each sentence
- Removing stop words

TextRank

TextRank is an iterative algorithm which uses a graph to rank units of text. The main steps of the using this algorithm are:

- A graph is built, where each vertex represents a sentence from the text
- Edges are formed between each two sentences with common words
- Weight is added to the edges based on the number of common words and the length of the sentences
- Sentences are ranked iterating the *TextRank* formula until convergence.

Extracting the sentences is done in two ways:

1. Taking the top predetermined percentage of sentences and sorting them by order in the original text to form a summary.
2. Using *UCS* to find a path between the first and the last sentence and form a summary from sentences on that path.

Noun rank

This algorithm builds a graph from nouns, ranks them and uses the ranks of the nouns to rank sentences in the text.

In order to maximize the number of nouns in the text, pronoun resolution is added to the preprocessing stage, and for identifying the nouns and pronouns, part of speech tagging is neccessery.

The main steps of using this algorithm are:

- Forming a graph, where each vertex represents a noun from the text.
- Adding an edge to the graph between each two nouns that are in a sentence together.
- Weight is added to each edge based on the distance of the nouns in their sentence
- Nouns are ranked based on the weight of the connected edges
- Sentences are ranked as a sum of the ranks of the nouns they contain
- The top ranked predetermined percentage of sentences is taken to form the summary

Results

- Running the algorithms on 216 Wikipedia articles and measuring time, compression rate and vocabulary retained for each algorithm gave the following results:

	TextRank	*TextRank UCS*	Noun rank
Average time	2.08s	1.64s	5.32s
Average compression rate	80%	91.66%	80%
Average vocabulary retained	22.83%	26.63%	28.12%

Conclusion

The three methods of text summarization vary significantly in different aspects of their performance. *TextRank* performs fast, but retains less vocabulary, indicating less information preserved. Using *UCS* to extract sentences results in very compressed summary. Noun rank retains the most information, but is also the slowest.