Introduction

“TextRank: Bringing Order into Texts” by Rada Mihalcea and Paul Tarau introduced TextRank, graph based, ranking algorithm for text processing.

In general, text based ranking algorithms are deciding on importance of a vertex of a graph by taking into account global information from entire graph. Connection from one vertex of a graph to another is based on voting. The higher number of votes the higher importance of vertex on whole graph. Algorithm takes initial, arbitrary value, iterates through each node in the graph, assign score to each vertex that assign importance of the vertex to whole graph. Final score is independent from initial, assigned value.

TextRank Algorithm

Mihalcea and Tarau created graph based ranking model applied to Natural Language Processing applications, contained unsupervised keyword extraction and sentence extraction. Graph based ranking algorithm used in natural language use weighted graphs, where we can define two vertices Vi and Vj as a weight wij that connects two vertices.

We need to not only represent the text but also connect words to other text entities in proper relations. This connections can be added to graph as a vertices, depending on application type and requirements.

Keyword extraction

Keyword extraction automatically identifies text or set of terms, that are the best description of whole text. Mihalcea and Tarau shows how their algorithm outperforms most of supervised algorithms, itself being unsupervised. They used units of text to create ranked summarization of whole text and represented the vertices. Co-occurrence relation was used to determine connection between vertices, and represented syntactic or semantic relations between elements of the text.

Algorithm works on tokenized text with part of speech tags. It takes all syntactic units of a given text, add it to the graph and set initial value of each vertex to 1. Then it iterates through the whole text until it converges. After obtaining scores algorithm sorts vertices in reversed order of its score and top T of scores are taken for post-processing, including marking potential keywords in text.

Evaluation of keyword extraction with TextRank was based on data set containing 500 abstracts from database Inspec. Each abstract also contains keywords extracted by professional indexers. Evaluation of algorithm was done using precision, recall and F-measure. Process was repeated using various filters, parameters and Co-occ.window value. In the best scenarios TextRank statistics was: 31.2 for Precision, 43.1 for Recall and 36.2 for F-measure. TextRank leaded to the highest F-measure compared to previously measured systems.

Sentence Extraction

Sentence extraction relies on keyword extraction approach. Instead of words or phrases we use entire sentences as a candidates for text units. Procedure for creating graph based representation of sentences is also quite similar to algorithm described, related to keyword extraction. Main difference is that now we have to rank whole sentence, therefore each sentence needs to represented as a vertex on a graph. There is also need for replacing co-occurance measure by similarity, being a function of content overlap of sentences, defined as a number of common tokens between representation of two sentences.

Evaluation of sentence extraction was proceeded using 567 news articles from DUC. TextRank was generating 100 words summary. The results was checked using Rogue evaluation toolkit. TextRank results was in top five best performances, but there was no parameters arrangement that would give TextRank the best score. The highest result 0.4904 was shown with stemmed text.

Summary

TextRank, in contrast to other tested algorithms is unsupervised and don’t require training corpora. After representing text data as a graph it can be successfully used for natural language processing.