Generic Text Summarization using Wordnet for Novelty and Hard

Ganesh Ramakrishnan, Kedar Bellare, Chirag Shah, Deepa Paranjpe
{hare,kedar,chirag,adeepa}@cse.iitb.ac.in
Dept. of Computer Science and Engg.,
Indian Institute of Technology, Mumbai, India

Abstract

This paper presents a Random Walk approach to text summarization using the Wordnet for text representation. For the HARD track, the specified corpus is indexed using a standard indexing engine - lucene and the initial passage set is retrieved by querying the index. The collection of passages is considered to be a document. In Novelty, the documents are as directly supplied by NIST. In either case, the document is used to extract a "relevant" sub-graph from the wordnet graph. Weights are assigned to each node of this sub-graph using a strategy similar to the Google Page-ranking algorithm. A matrix of sentences against the nodes of the sub-graph is created and principal component analysis is performed to extract the sentences for the summary. Our approach is not specific to any particular genre of documents, such as news articles. We use the semantics in the document rather than using the more common statistical measures like term frequency and inverse-document frequency.

Keywords : text summarization, wordnet, PCA, synsetranking

1 Introduction

Text summarization is "the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)". Text summarization finds varied applications in today's world. Some notable ones are: search engine hit summarization (summarizing the information in a hit list retrieved by some search engine); physicians' aids (to summarize and compare the recommended treatments for a patient); generating the blurb of a book ; and so on. Building automated summarizers can be very helpful in many such applications and saves a lot of manual work. An extract is a summary containing only material from the text and involves no natural language generation. Given a piece of text, our aim is to select the most "representative" sentences which will form the summary.

A good summary should ideally have the following features

1. Relevance to the text

2. Informativeness

3. Conciseness

1.1 Our approach

For the HARD track, we index the document collection using the lucene indexer. The query is fired to the search component of lucene and a set of relevant passages are extracted after lucene querying. These passages are combined to form a document which is further subject to summarization. In Novelty, the documents are as directly supplied by NIST.

We use wordnet to understand the links between different parts of the document; Subsequently, we extract the portion of the wordnet graph which is most relevant and contains the main ideas present in the document. We do this by starting from the words in the document as the leaves and traverse the wordnet links upward towards synsets that are more general at each level. The idea of first getting a global view of the whole document, even before beginning to rank sentences is what differentiates our approach from the rest. After obtaining the relevant sub-graph we rank its nodes (synsets of wordnet) with random walks on wordnet and use them to compute the importance of individual sentences.

This idea of getting an overall picture of the document before picking sentences makes our approach generic (not necessarily tailored to give good results only on a specific class of documents).

The last, and the most crucial phase of our approach is to actually pick out sentences that will form the summary. It is important that we do not pick two sentences with a similar meaning and only pick those sentences which represent the text to the maximum extent possible. For this purpose, we use the method of **Principal Component Analysis** and dump sentences most higly correlated with each principal component as the summary sentences. This is where, we believe, our approach captures human thinking. It is natural for a human to identify very similar sentences from the text and pick only one of them.

2 Conclusion

There is an ever-increasing need for better automatic text summarization systems with the explosion in the amount of information available the user. Most existing text summarization systems analyze a text statistically and linguistically, determine important sentences, and generate an extract for it. The linguistic features used are generally specific for a particular kind of document which make existing systems very specialized. We propose an algorithm for a generic text summarizer which selects sentences on the basis of their semantic content and its relevance to the main ideas contained in the text. We use Wordnet to abstract the ideas contained in the text so that sentences are selected on the basis of their meaning, and not on the presence of some keywords or frequently-occuring term.