Relevant Information Retrieval from Text Documents

Extracting relevant information from a document is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points.

There are two methods

1. Extractive information retrieval
2. Abstractive information retrieval

Extractive methods work by identifying important sections of the text and generating them verbatim.

Abstractive methods aim at producing important material in a new way. In other words, they interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information from the original text.

1. Extractive Information Retrieval

In order to better understand how system work, we describe three fairly independent tasks

- Construct an intermediate representation of the input text which expresses the main aspects of the text. There are two types of approaches based on the representation: topic representation and indicator representation. Topic representation-based IR techniques differ in terms of their complexity and representation model, and are divided into frequency-driven approaches, topic word approaches,
  latent semantic analysis and Bayesian topic models
- Score the sentences based on the representation.
- select a summary comprising of a number of sentences

2. Abstractive Information Retrieval

Abstractive IR does not simply copy important phrases from the source text but also potentially come up with new phrases that are relevant.

3. System Evaluation

   ROUGE is the most widely used metric for automatic evaluation.