Approaches to Language Understanding for Information Retrieval

Gerard Salton
Department of Computer Science
Cornell University

The effectiveness of text processing applications such as information retrieval depends in large part on the adequacy of the content analysis of the texts and the representation of the text meaning. The interpretation of natural-language text is unfortunately complex and is substantially controlled by outside factors such as the circumstances and contexts in which the text utterances appear, and the background and experiences of writer and reader. In these circumstances, the text understanding problem might be tackled by undertaking a detailed analysis of large bodies of text in the hope of determining text contexts where word meanings are relatively unambiguous and text interpretation becomes possible.

Methods are covered for breaking down large text samples into smaller local documents, and for clustering and linking local documents of varying size. The attempt is to produce representations of large texts as linked structures of local items in such a way that text meanings are unambiguous in each local substructure. Applications of these techniques are described in automatic information retrieval.

[handwritten notes:]

* L. Wittgenstein - Philosophical Investigations

- Treat text of heading differently than paragraph text

- informal vs formal message
  → spelling