Aspects of Sentence Retrieval

Vanessa Murdock[1]
Center for Intelligent Information Retrieval
University of Massachusetts, Amherst
vmurdock@yahoo-inc.com

Sentence Retrieval is the task of retrieving a relevant sentence in response to a query, a question, or a reference sentence. Tasks such as question answering, summarization, novelty detection, and information provenance make use of a sentence-retrieval module as a preprocessing step. The performance of these systems is dependent on the quality of the sentence-retrieval module. Other tasks such as information extraction and machine translation operate on sentences, either using them as training data, or as the unit of input or output (or both), and may benefit from sentence retrieval to build a training corpus, or as a post-processing step.

In this thesis we begin by demonstrating that because sentences are much smaller than documents, the performance of typical document retrieval systems on the retrieval of sentences is significantly worse. We propose several solutions to the problem of sentence retrieval, and investigate these solutions the application areas of sentence retrieval for question answering, novelty detection, and information provenance.

The context of a sentence affects its meaning, and we demonstrate that smoothing from the local context of the sentence improves retrieval when the collection to be retrieved from contains many documents of unknown relevance.

We show that statistical translation models are appropriate for tasks where the sentence to be retrieved has many terms in common with the query, but still benefits from the addition of related terms and synonyms. We show that queries of very few terms benefit from the translation approach, which incorporates related terms into the query. We show that the family of language modeling approaches, which includes statistical translation models, is not effective for discriminating between sentences that use the same vocabulary to express the same information, and sentences that use the same vocabulary to express new information. Finally, we demonstrate a conditional model for sentence retrieval for question answering, and show that it outperforms both the translation approaches and the baseline language-modeling approach.

[1] Author's current affiliation: Yahoo! Research Barcelona