Utilizing Sub-topical Structure of Documents for Information Retrieval

Debasis Ganguly
CNGL, School of Computing,
Dublin City University, Ireland
dganguly@computing.dcu.ie

Recent years have witnessed an upsurge in the quantity of news, encyclopedic articles, blogs, forum and social networking posts etc. over the web. Some of these, such as the news and Wikipedia articles are carefully authored, edited and quality controlled, while others such as blogs and social networking posts are not. A document in the former category is often explicitly decomposed into paragraphs or sections to convey a specific aspect of the overall information in a more focused way. The sub-topic based organizational pattern is more implicit in a document of the later category, due to the absence of explicitly demarcated paragraphs. It is very important to gain insights into the sub-topical structure of documents to effectively manage and utilize the different aspects of information contained therein. The process of mining out the sub-topical structure from a document is known as text segmentation.

Our work centers around applying text segmentation for a multitude of information retrieval (IR) tasks such as scoring a document combining the retrieval scores of its constituent segments, exploiting the proximity of query terms in documents for ad-hoc search, and for question answering (QA), where retrieved passages from multiple documents are aggregated and presented as a single document to a searcher.

Feedback in ad-hoc IR task is shown to be benefited by the use of extracted sentences instead of terms from the pseudo relevant documents for query expansion. Retrieval effectiveness for patent prior art search task is enhanced by applying text segmentation to the patent queries by two methods. In one method, we decompose a patent query into sentences or fixed word length windows, accumulate the similarities of every segment over the set of pseudo relevant documents, and remove the ones which are least similar thus ensuring that the reduced query to be used in the feedback step is more similar to the pseudo relevant set of documents. This method of query reduction increases MAP by 7.28% relative to the second best performing system of CLEF-IP 2010 which involved a naive method of query reduction by removal of unit frequency terms. The second method uses text segmentation to decompose a patent query description into constituent sub-topics and uses each sub-topic as a separate query for retrieval and finally merges the results from these different query streams. Decomposing a patent query this way, is shown not only to improve the performance of initial retrieval but also to enhance feedback effectiveness. The results show that query segmentation alone can result in a 12.14% increase in PRES and segmentation coupled with query expansion results in a 14.05% increment of PRES.

We are also exploring on augmentation of text segmentation techniques to produce more readable segments with less unresolved anaphora. This particularly is useful for QA and snippet generation tasks where the objective is to aggregate relevant and novel information from multiple documents satisfying user information need on one hand, and ensuring that the automatically generated content presented to the user is easily readable on the other.

Text segmentation, Pseudo relevance feedback, Question Answering