BlogSum: A Query-based Summarization Approach to Make Sense of Social Media

Shamima Mithun
Computer Information Technology Department
IUPUI

With the rapid growth of the Social Web, a large amount of informal opinionated texts are available on numerous topics. However, people can be overwhelmed with this vast amount of information and they need help to find the information of their interests. Natural language tools for automatically analyzing these opinions become necessary to help individuals, organizations, and governments in making timely decisions. To address this need, I proposed a summarization approach for opinionated texts. To validate my approach, BlogSum is developed and evaluated experimentally using current benchmarks. Users can ask BlogSum any question (e.g. Why do people like Chrome better than Firefox?). To answer user's question, BlogSum first retrieves relevant blogs, reviews from the web then generates a concise summary that represents people opinions expressed towards the topic.

Since blog summarization is a more recent endeavor, an error analysis was conducted by manually analyzing blog summaries to find there is any information processing difference needed for blogs compared to factual data. This analysis shows that question irrelevance and discourse incoherence, which decrease the overall quality of a summary and reduces the summary coherence, are two major issues for blog summaries. To address question irrelevance and discourse incoherence, in this work a domain-independent schema-based summarization approach is developed that utilizes discourse structures. This approach is based on the automatic identification of discourse relations within candidate sentences in order to instantiate the most appropriate discourse schema and filter and order candidate sentences in the most effective way. BlogSum also needs to deal with opinions, emotions effectively to be successful. BlogSum's overall performance as well as performance for question relevance and coherence was evaluated using various dataset. These results show that the proposed approach can effectively reduce question irrelevance and discourse incoherence and satisfy user's information need.