Entity-Centric Summarization

Shruti Chhabra Srikanta Bedathur
Indraprastha Institute of Information Technology, New Delhi
{shrutic,bedathur}@iiitd.ac.in

ABSTRACT

Web search has become ubiquitous and with the availability of huge volumes of data on web and effective retrieval systems, user can now obtain good-quality information relevant to his need easily. But, still, finding comprehensive information about an entity is a laborious task and requires efforts by user to gather diverse and relevant information from the multiple sources and form a summary. In our research, we propose a framework to tackle the problem of automatically generating a single document summary for ad-hoc topic (entity) posed by user. Relationships between the input entity and associated entities are identified to obtain comprehensive information about the entity.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Miscellaneous

General Terms

Information Retrieval

Keywords

summarization, related entity finding, diversity

1. INTRODUCTION

Web is now pervasively searched by users to satisfy their information needs. A search on the Web presents the user with the most relevant documents, as judged by the system, for the posed query. Unfortunately, in many cases, the required information is scattered. As a result, user may not have enough time to read all the documents and therefore, some information may be left unread.
Also, with the tremendously increasing information on web, it is getting difficult for the user to gather information from multiple sources for the queried topic and extract the relevant and diverse information to form a digest of the topic.

I-CARE 2012

Although, highly rich information source Wikipedia [1] is present. It is a human-generated encyclopedia which provide comprehensive content about various entities but still it's not exhaustive. On posing the query "1971 war" to a search over Wikipedia, it can immediately direct us to a page about "Indo-Pakistani War of 1971" but less popular entities and events like "Shruti Chhabra" still remain unrepresented.
Entity-centric summarization assists the user by presenting a single document containing relevant information about the entity searched. Sentences in the summary should be coherently ordered and each sentence should convey non-overlapping information (diversified sentences). Entity-centric summarization can be classified as topic-driven multi-document summarization task, but, herein, the topic ought to be an entity. This task is challenging due to various factors. An entity can have various surface forms example "Mark Zuckerberg" can be mentioned as "Mark E. Zuckerberg","Mark Elliot Zuckerberg" or "Facebook Creator". Entities are usually mentioned in documents in their various surface forms. Hence, it pose the challenge of considering all the mentions of entity, during sentence retrieval for summary creation.
In literature, the problem has been approached from various directions. Sauper et. al. [3] used high-level structure of human-generated text to automatically create domain specific templates. Topic-specific extractors are learned jointly for the entire template and used for content selection. They work only on Disease and American Film Actors entities, which exhibit fairly consistent article structures hence facilitating a good quality template. Filippova et. al. [2] built a multi-document summarization system to obtain company-specific summaries from financial news. Sentences are selected on the based on relatedness to the company symbol (expanded with related words) , overall importance and novelty. They use "business summaries" of Yahoo! Finance to expand their input query. Hence, these approaches limit to domain-specific queries.
In this research work, we propose a framework to automatically generate summaries for input entity. Every entity is generally associated with various other entities. For Example, entity "Manmohan Singh" is associated with "India", "Prime Minister", "Sonia Gandhi", etc. A lot of research has already been done in this field. Various commercial applications on web like Evri, TextMap, ZoomInfo, allow users to search and browse entities related to a topic or to another entity.

[1] http://www.wikipedia.org/

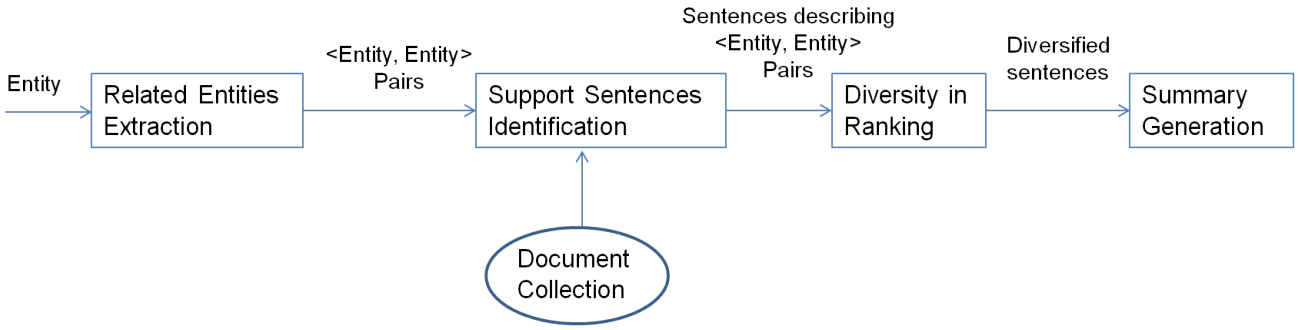

Figure 1: Proposed Framework

Our approach is based on identifying sentences depicting all such important entity pair relationships. These sentences, then, need to be ranked and diversified to form a summary of the topic.

2. PROPOSED FRAMEWORK

The proposed framework is based on identifying sentences which describe relationship between the input entity and its associated entities as shown in Figure 1.

In the framework, four modules are pipelined : Related Entities Extraction which on inputting an entity, finds the entities associated to the entity, Support Sentence identification which identifies the sentences describing the relationship between given entity pair and Diversity in Ranking module which eliminates the redundant information in the support sentences.

The concept of finding sentences describing relationship between a query and associated entities was first formalized by Blanco et. al. [1]. They proposed ranking model based on bag-of-model approach, positional measure and entity ranking based approach for these sentences. On empirical evaluation of these methods, they found that context (i.e. two preceding sentences, two succeeding sentences and the title of document in their case) plays an important role in enhancing the performance of the sentence retrieval. We incorporate the concept proposed by them, in our work.

The input entity, say e, is fed to the related entities extraction module which outputs all the important entities which are related to e. Let us call the obtained entities as $e_1, e_2, ..., e_n$. The entities are expanded with their surface forms, and called $e', e'_1, e'_2, ..., e'_n$. These surface forms are extracted from a dictionary built using Wikipedia structure (Redirect pages, Disambiguation pages, Hyperlinks). Then, the sentences describing the relationship between the (expanded) entity pairs $< e', e'_1 >, < e', e'_2 >, ..., < e', e'_n >$ are identified. Such sentences are referred as "Support Sentences". Context of the sentence is considered while ranking the sentences; this context can be defined as surrounding sentences, a passage, the document's title or even the entire document. The ranking model is defined as:

$$H_{ee'}(s) = F_e(s, C_s) \qquad \forall\ s \in \mathbf{S}_{ee'} \tag{1}$$

where C_s is the context of the sentence s, $\mathbf{S}_{ee'}$ is the sentence mentioning both the entities e and e' and function F_e refers to any field sensitive ranking model like BM25F, ranked for the entity e.

The identified ranked support sentences may contain overlapping information. The redundant information from the sentences needs to be removed. So, these sentences are passed through the Diversity in Ranking module. The sentences are re-ranked, so that as we go down the ranked list of sentences, each sentence adds substantial value to the information conveyed so far. These re-ranked sentences when concatenated form a comprehensive summary of the entity inputted.

3. EXPECTED RESULTS

In real world, any entity is defined based on the entities it is connected to and the relationships it holds with them. Today's web structure closely maps the entire real world scenario. Based on the same intuition, we expect that the proposed framework will be capable of capturing the overall overview of the entity. Also, since every entity follows the same pattern, we expect the system to be domain-independent.

4. CONCLUSION

In this research work, we have proposed a framework to form a comprehensive summary about the input entity. Through this approach, we try to explore the new generalizable idea of capturing relationships between entities for producing summaries.

5. REFERENCES

[1] Blanco, R., and Zaragoza, H. Finding Support Sentences for Entities. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), pp. 339–346.

[2] Filippova, K., Surdeanu, M., Ciaramita, M., and Zaragoza, H. Company-oriented Extractive Summarization of Financial News. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (2009), pp. 246–254.

[3] Sauper, C., and Barzilay, R. Automatically generating wikipedia articles: a structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1* (2009), pp. 208–216.