Generating Text Summaries of Graph Snippets

Shruti Chhabra
Indraprastha Institute of Information Technology
New Delhi, India
shrutic@iiitd.ac.in

Srikanta Bedathur
Indraprastha Institute of Information Technology
New Delhi, India
bedathur@iiitd.ac.in

ABSTRACT

With the availability of large entity-relationship graphs, finding the best relationship between entities is a problem that has attracted a lot of attention. Given two or more entities, the goal of most algorithms is to produce a graph structure of varying complexity (i.e., a simple path, a minimal weighted tree, or a dense subgraph etc.) as a way of characterizing the relationship between given entities. However, no attention is paid to the interpretability of these results – i.e., the ability of humans to read these and comprehend the context in which these relationships exist. A key obstacle in this direction is the lack of necessary linguistic context and natural textual result formulations.

We pursue the idea of using *entity-centric summarization* as a way of closing this gap. We aim to turn the resulting graph structures into one or more coherent textual snippets (or summaries) that can be easily read and interpreted. In this short position paper, we first outline two different scenarios that result in slightly different formulations of the problem. Based on preliminary experimental results, we discuss the challenges that are inherent in this setting.

1. INTRODUCTION

Recent years have seen an explosive growth in the area of automatic information extraction from large text corpora ranging from the entire Wikipedia to online news articles. As a consequence, we have access to big, automatically populated knowledge graphs such as DBPedia [2], FreeBase [4], Yago2 [18] etc. On such graphs, a natural question to ask is *"Given a set of entities, how to best characterize the relationships between these entities?"*. Many recent results addressing this question have the following structure: consider the graph structure of the knowledge graph (possibly with the type/relationship information), and compute a smaller/simpler structure that best describes the common relationship(s) among the input entities. The resulting structure could be a simple path [8], a minimum weighted Steiner tree [12], or a general subgraph that connects the input entities [7] by optimizing certain graph-structural property such as sum of edge-weights, density of edges, conductance, etc. Unfortunately, none of these solutions specifically pay attention to the interpretability of results. In many cases the resulting graph structure is either very hard to interpret, or requires extensive background and contextual information to do so.

On the other hand, the text sources that were used for populating these knowledge graphs contain much more information than just what has been extracted. With this insight, many researchers have considered the problem of generating so called *support documents* [3] that helps in the better understanding of a relationship. The user is burdened with the task of combining these support documents for relationships in the result graph.

In this work, we aim to overcome this limitation and make the relationship graphs interpretable with relative ease by end users, and at the same time, retain the power of graph analytics based knowledge discovery which goes beyond the document-level search. The idea we pursue is that of generating *entity-centric summaries* for one or more entities connected in the form of a simple graph structure. Specifically, in this paper we formulate two different problems that correspond to two different graph structures: first, *Entity summaries* that take a single central entity and generate a textual summary by considering all the entity-relationships that the input entity participates in (i.e., a star graph around the entity), and second, *chain summaries* that, as the name suggests, take a chain of relationships that connect two entities and compute a summary that combines text snippets of each relationship involved in the path.

2. ENTITY-CENTRIC SUMMARIES

In this section, we present the two variants of entity-centric text summaries that we consider in this paper and develop problem formulation. The first variant we consider is called *entity summary* which considers one input entity along with all the relations that it is involved in, and uses them along with text snippets associated with each relationship to generate a combined text summary of the entity. In the next one, which we call as *chain summaries*, a relationship chain connecting two entities is given as input, and the goal is to compute a ranking of text summaries which describe the relationship path. Before describing these two variations, we will establish the model on which we build our solutions.

2.1 Model

Our overall framework consists of a knowledge graph

The 19th International Conference on Management of Data (COMAD), 19th-21st Dec, 2013 at Ahmedabad, India.

Figure 1: Example of a star graph of query entity.

G = (V, E) consisting of a set of entities V and *labeled relationships* between them represented by the set of edges E. Although many knowledge-bases represent relationships as directed edges, we will ignore the directionality for simplifying our model. In addition to labels, each relationship can also be seen to contain a ranked list of sentences from the text corpus from which the entity-relationships were extracted. In fact, many modern knowledge-bases, such as NELL [**?**], Open IE [6] and Yago2 [18], explicitly maintain the *evidence sentences* from which the relationships were extracted. In absence of such information, it may be possible to retrieve short passages or sentences from the text corpus through techniques such as support sentence retrieval [3].

We represent the set of text snippets associated with each edge $e \in E$ as S_e. Now, if we consider two edges e_1 and e_2 that have a common node, and corresponding sentence sets S_{e_1} and S_{e_2}, then the goal is to obtain a partial ordering of

$$l_1 \oplus l_2, \quad \forall l_1 \in S_{e_1}, \forall l_2 \in S_{e_2},$$

where $\oplus$ indicates concatenation of the two snippets l_1 and l_2.

2.2 Entity Summaries

Consider the graph shown in Figure 1 illustrating the relationships between various entities with Leonardo Da Vinci as the central entity. When one needs to get an overview of life and works of Da Vinci, the entire set of relationships that the entity Leonardo Da Vinci is involved in contains all the necessary information. A descriptive text for the central entity (in this case Leonardo Da Vinci) can be generated by fusing the text segments of these relationships.

This problem has significant overlaps with a few other directions of research addressed in the recent past. Liu *et al.* [13] considered web block as the basic coherent unit for their task and proposed an integrated bootstrapping model BioSnowball to jointly solve the biography ranking and fact extraction to summarize the Web to generate Wikipedia-style pages for any *person*. On the other hand, Sauper *et al.* [17] 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 consider *Disease* and *American Film Actors* entities, which exhibit fairly consistent article structures hence facilitating a good quality template. Filippova *et al.* [9] built a multi-document summarization system to obtain *company specific* summaries from financial news articles.

2.3 Chain Summaries

In many entity analytics tasks, one of the typical queries is to find a chain of relationships to connect two query entities. Given such a chain, represented as $\langle e_1, e_2, ..., e_n \rangle$, we aim to produce a partial ordering (i.e., ranking) of concatenation of sentences,

$$l_1 \oplus l_2 \oplus ... \oplus l_n, \quad l_1 \in S_{e_1}, l_2 \in S_{e_2}, \ldots, l_n \in S_{e_n}.$$

Generating textual relationships for chain of entities is a relatively less explored research problem. Jin *et al.* [11] aimed at finding most meaningful evidence trails across documents that connect topics. The technique is tested for a specific dataset, therefore, the effectiveness for generic queries is unclear. We propose to utilize the entity chain derived from entity graph and break the graph into entity pairs. We retrieve sentences for each entity pair and concatenate them to form summary.

3. FRAMEWORK

The framework for our automated entity-centric summarization system is composed of three modules (See Figure 2) in following sequence:

1. Graph Construction - Construction of Star Graph or Entity Chain based on the input type - entity or pair of entities.
2. Identifying Relationship Descriptions - Selection of candidate sentences which may describe the relationship (edge) between entities.
3. Coherent Sentence Sequencing - Align the resultant sentences from previous module to create a coherent summary.

Each module in the framework is in itself an open research problem. We will now discuss each module in the framework and inherent challenges.

3.1 Graph Construction

Based on the input, either star graph or entity chain would be constructed as explained in this section.

- If the input is an entity, the entities related to the input entity are retrieved and linked in a star fashion.
  Finding Related Entities : *Given an entity e, the task is to find top-k entities $e_1, e_2, ..., e_k$ related to e and cluster these entities based on the relation type with e.*
  Finding related entities is a well researched problem. TREC[1] 2009, 2010 and 2011 introduced related entity finding (REF) task through its entity track, however, the dataset provided a lot of information and context about the target entities, which is generally not available in real-time. There are knowledge bases *derived from Wikipedia* like Yago, which provide huge data on triples (facts) of entity relationships. Also, there have been efforts, like NELL [5], to create an ontology from unstructured web pages.
  Apart from finding related entities, ranking these entities is also a complex and challenging problem. Related entities can be ranked based on their relationships with the central entity. Therefore, to find the top-k related entities, the strength of relationships needs to be quantized. Entities can be related to the central entity in various contexts such as "personal", "political", "social". Therefore, context-based clustering of related

[1]http://trec.nist.gov/data/entity.html

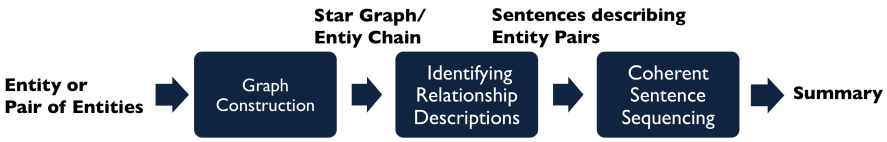

Figure 2: Proposed Framework

entities is required to facilitate the process of coherent sequencing of sentences at later stages.

- If the input is a pair of entities, entity chain is constructed.
  Forming Entity Chains : *Given a entity pair e_1 and e_2, the goal is to obtain a chain of entities $e_1, e_2,..., e_k$ through which e_1 and e_2 are connected.*
  Researchers from various domains have attempted to address the problem of extracting transitive relationship between entities. In biomedical domain, text mining algorithms, commonly termed as *Hypothesis Generation*, are designed to extract novel associations (MeSH terms/ concepts) from the unstructured or semi-structured biomedical text. The text mining techniques have also been applied on web collections to find evidence sentences across documents that connect two input topics either directly or through intermediate topics. This special case of text mining is referred as *Concept Chain Queries* [11]. There have been efforts to find such semantic associations from RDF [1] and create clique chains between entities [10].

3.2 Identifying Relationship Descriptions

Given an edge e_i, the task is to find ranked list of relationship descriptions, S_{e_i}.

The task of extracting relationship descriptions appears similar to a sentence retrieval problem. Sentence Retrieval is the task of retrieving a relevant sentence in response to a query, a question, or a reference sentence [15]. But, unlike sentence retrieval problem, our problem deals with pair of entities as query and aims at retrieving sentences describing relation between both. The traditional sentence retrieval techniques cannot retrieve sentences for a pair of entities efficiently as there are additional technical issues such as defining appropriate ranking measures, considering context, etc., to be addressed. An entity can also be present in various surface forms which makes the problem more complex. For instance, "Mark Zuckerberg" can be mentioned as "Mark E. Zuckerberg","Mark Elliot Zuckerberg" or "Facebook Creator". Hence, it is important to consider all the occurrences of the entity during sentence retrieval.
Unlike our notion of relationship as a sentence, researchers have represented relationships as Connection Subgraphs [7], Steiner Tree [12], Semantic Associations [1], Relationship Explanations [8], Clique Chains [10]. However, Blanco *et al.* [3] describe the direct relationship between the query and associated entities as sentences. They termed such sentences as "Support Sentences".
Various tools such as Reverb[2], Open IE [6], Patty [16], have been introduced which output the sentences describing the relationship between the two input entities.

3.3 Coherent Sentence Sequencing

[2]http://reverb.cs.washington.edu

Given all re-ranked sentence sets $S_{e_1}, S_{e_2}, ..., S_{e_n}$, the task is to produce following partial ordering

$$l_1 \oplus l_2 \oplus ... \oplus l_n, \quad \forall l_1 \in S_{e_1}, \forall l_2 \in S_{e_2}, ...\forall l_n \in S_{e_n}$$

Coherence is a property of well-written texts that makes them easier to read and understand. The text coherence should be considered at two levels: 1) local coherence which implies sentence to sentence transitions should be smooth, 2) global coherence which considers discourse-level relation connecting remote sentences. The coherence can be modeled in the guise of topical closeness, lexical coherence, temporal coherence, content relatedness, etc. Sentence ordering also plays a vital role to capture text coherence.
A number of different theories from a variety of intellectual disciplines have been proposed to represent coherence in multi-sentence text, including RST, Discourse Grammar, Macrostructures, Coherence Relations, etc [14]. We need to consider relevance scores of each sentence, in addition to local and global coherence measures so that coherence score of combination of highly relevant sentences is higher.

4. EXPERIMENTS

In this section, we discuss our preliminary experimental evaluation. The purpose is to analyze the results obtained by clubbing the existing techniques for different modules and identify the various associated challenges.
In our experiment, we assume that we already have the graph (star graph or entity chain graph) and aim to find the summary of the graph. The query set comprises of varied types of entities such as person, organization, date, country, product.
We utilize Open IE [6] to identify set of relationship descriptions for edges between entity pairs. Open IE extracts basic relations from web corpora and allows search over the learned relations. It provides various relationship keyphrases and associated sentences (relationship descriptions) for any subset of {argument1, verb phrase, argument2}.
In the case of an entity chain graph, the identified relationship descriptions are concatenated based on the cosine similarity distance measure (a lexical coherence measure). The most similar sentences in the set of relationship descriptions for first entity pair are concatenated. For other entity pairs, sentence is selected which is most similar to the sentence selected for the previous adjacent pair. On the other hand, for star graph, keyphrase with maximum number of sentences is identified. The top sentence from the sentence set associated with the key phrase is thus selected. Summary is obtained by fusing the selected sentences. Results for few example queries are summarized in Tables 1 and 2.

5. DISCUSSION

In this research, we propose a novel framework to address problem of entity-centric summarization. The research questions and technical challenges associated with the framework

Table 1: Results for sample Entity Chain Graphs

Graph Edges	Top Concatenated Relationship Description	Better Concatenated Relationship Description (rank)	Remarks / Insights
(Tim Cook, Apple Inc., Steve Jobs)	When Tim Cook took over the helm at Apple Inc. in August , many wondered how he would set himself apart from his predecessor , the formidable Steve Jobs . Friday , March 30th , 2012 at 9:15 AM When the late Steve Jobs stepped down from Apple in 2011 , many wondered how well his successor , current CEO Tim Cook would be received by Apple employees .	The top sentence captures the relationships in the entity chain.	Even a single sentence is able to describe relation among all entities. Redundancy needs to be minimized. Ranking based on coherence needs to include relevance score of relationship descriptions with respective entity pairs. Non-relevant information present should be removed. Based on the type of connection (causative, elaborative, etc), connectors are missing. The level of difficulty increases when entities are associated with each other through various types of relations such as personal, political. E.g. Monica Lewinsky and Bill Clinton.
(Monsanto, Bt Cotton, Bacillus Thuringiensis)	Then in 1996 , Monsanto introduced BT cotton - a GMO that employs a gene from the bacterium Bacillus thuringiensis to make a powerful pesticide in the plant . Bt cotton , widely grown around the world , contains a gene from Bacillus thuringiensis , a bacterium species .	The top sentence captures the relationships in the entity chain.	
(Iran, Saddam Hussein, Iraq)	Iran is the enemy of Arabs , Islam and the United States , and the only person who can stand in the face of Iran is Saddam Hussein , he said . The desired result of this planned war on Iraq is the overthrow of Saddam Hussein , the destruction of all chemical , biological , and nuclear weapons facilities , and the maintenance of Israel's position as the most powerful nation (and the only nuclear power) in the region .	Hussein declared war on Iran in 1980 , he had the understandable backing of the United States , but the war ground to a stalemate in 1988 after costing both nations hundreds of thousands of casualties . When Saddam Hussein led Iraq in the seizure of Kuwait in August of 1990 , the United States took advantage of his exaggerated ambitions to make war on Iraq . (6)	
(Hillary Rodham Clinton, Bill Clinton, Monica Lewinsky, Linda Tripp)	Mr. Clinton grabbed Hillary by the scruff of the neck and heaved her over the railing. Bill Clinton called Monica Lewinsky at 2:00 AM the next morning to tell her of the death and alert Monica that her named appeared on the Paula Jones witness list . Unknown to either Lewinsky or President Clinton , a former White House staffer Linda Tripp had recorded Lewinsky talking about her relations with President Clinton .	Hillary is married to Bill , and Bill will come to the WH with her . News in the US is dominated by the allegation that Bill Clinton had a sexual relationship with Monica Lewinsky and persuaded her to lie under oath . Unknown to either Lewinsky or President Clinton , a former White House staffer Linda Tripp had recorded Lewinsky talking about her relations with President Clinton . (2)	

Table 2: Results for sample Star Graph

Graph Edges	Relationship Description	Remarks / Insights
(Leonardo da Vinci, 1452)	Leonardo da Vinci was born in 1452, a hundred years after the exhibition.	Sentence Selection for an edge will depend on other edges. Sentence ordering is important. Temporal aspect needs to be considered. Related entities need to be clustered based on the type of relation.
(Leonardo da Vinci, Italy)	Let 's trace the history of this rumor ... Following the death of his patron Giuliano deMedici , Leonardo left Italy and came to France in 1516 , at the age of 64.	
(Leonardo da Vinci, Mona Lisa)	Leonardo Da Vinci painted The Mona Lisa, Many have flocked to The Louvre to marvel at its beauty.	
(Leonardo da Vinci, France)	Leonardo died in France and is buried in the grounds of the chateau .	
(Leonardo da Vinci, Amboise)	—	
(Leonardo da Vinci, Bill Gates)	The record sale price of a nonliterary manuscript was $ 30.8 million that Microsoft chairman Bill Gates paid for Leonardo da Vinci 's 16th-century scientific tract , the Codex Leicester , in 1994.	

are summarized below:

- How to cluster the related entities obtained? Should each cluster be summarized separately? Can clustering be more effective at the time of finding coherent sentence sequences?
- How to quantize strength of relationships and thus rank related entities?
- What should be the ranking model to find sentences describing relationship between entities?
- How to combine relevance and coherence to obtain the partial ordering?

6. REFERENCES

[1] K. Anyanwu and A. Sheth. The ρ operator: discovering and ranking associations on the semantic web. *SIGMOD Record*, 2002.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 2009.

[3] R. Blanco and H. Zaragoza. Finding Support Sentences for Entities. SIGIR, 2010.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. SIGMOD, 2008.

[5] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[6] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam. Open information extraction: the second generation. IJCAI, 2011.

[7] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. KDD, 2004.

[8] L. Fang, A. D. Sarma, C. Yu, and P. Bohannon. Rex: Explaining relationships between entity pairs. *VLDB Endowment*, 2011.

[9] K. Filippova, M. Surdeanu, M. Ciaramita, and H. Zaragoza. Company-oriented Extractive Summarization of Financial News. EACL, 2009.

[10] M. S. Hossain, P. Butler, A. P. Boedihardjo, and N. Ramakrishnan. Storytelling in entity networks to support intelligence analysts. KDD, 2012.

[11] W. Jin, R. K. Srihari, H. H. Ho, and X. Wu. Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. ICDM, 2007.

[12] G. Kasneci, M. Ramanath, M. Sozio, F. M. Suchanek, and G. Weikum. Star: Steiner-tree approximation in relationship graphs. ICDM, 2009.

[13] X. Liu, Z. Nie, N. Yu, and J.-R. Wen. Biosnowball: automated population of wikis. KDD, 2010.

[14] I. Mani. *Automatic summarization*, volume 3. John Benjamins Publishing, 2001.

[15] V. Murdock. Aspects of sentence retrieval. *SIGIR Forum*, 2007.

[16] N. Nakashole, G. Weikum, and F. Suchanek. Patty: a taxonomy of relational patterns with semantic types. EMNLP-CoNLL, 2012.

[17] C. Sauper and R. Barzilay. Automatically generating wikipedia articles: a structure-aware approach. ACL-IJCNLP, 2009.

[18] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. WWW, 2007.