Visualizing Multi-Document Semantics via Open Domain Information Extraction

Yongpan Sheng[1], Zenglin Xu[1], Yafang Wang[2], Xiangyu Zhang[1], Jia Jia[2], Zhonghui You[1], Gerard de Melo[3]

[1]School of Computer Science and Engineering,
University of Electronic Science and Technology of China, China;
[2]Shandong University, China; [3]Rutgers University, USA
{shengyp2011,zenglin,wyf181,keposmile.z,jiajia911,zhyouns}@gmail.com
gdm@demelo.org

Abstract. Faced with the overwhelming amounts of data in the 24/7 stream of new articles appearing online, it is often helpful to consider only the key entities and concepts and their relationships. This is challenging, as relevant connections may be spread across a number of disparate articles and sources. In this paper, we present a system that extracts salient entities, concepts, and their relationships from a set of related documents, discovers connections within and across them, and presents the resulting information in a graph-based visualization. We rely on a series of natural language processing methods, including open-domain information extraction, a special filtering method to maintain only meaningful relationships, and a heuristic to form graphs with a high coverage rate of topic entities and concepts. Our graph visualization then allows users to explore these connections. In our experiments, we rely on a large collection of news crawled from the Web and show how connections within this data can be explored.

Keywords: Multi-Document information extraction, Graph-based visualization.

1 Introduction

In today's interconnected world, there is an endless 24/7 stream of new articles appearing online, including news reports, business transactions, digital media, etc. Faced with these overwhelming amounts of information, it is helpful to consider only the key entities and concepts and their relationships. Often, these are spread across a number of disparate articles and sources. Not only do different outlets often cover different aspects of a story. Typically, new information only becomes available over time, so new articles in a developing story need to be connected to previous ones, or to historic documents providing relevant background information.

In this paper, we present a system that extracts salient entities, concepts, and their relationships from a set of related documents, discovers connections within and across them, and presents the resulting information in a graph-based visualization. Such a system is useful for anyone wishing to drill down into datasets and explore relationships, e.g. analysts and journalists. We rely on a series of natural language processing methods, including open-domain information extraction and coreference resolution, to

achieve this while accounting for linguistic phenomena. While previous work on open information extraction has extracted large numbers of subject-predicate-object triples, our method attempts to maintain only those that are most likely to correspond to meaningful relationships. Applying our method within and across multiple documents, we obtain a large conceptual graph. The resulting graph can be filtered such that only the most salient connections are maintained. Our graph visualization then allows users to explore these connections. We show how groups of documents can be selected and showcase interesting new connections that can be explored using our system.

2 Approach and Implementation

2.1 Input Data

Our system is designed to operate on a large collection of news articles. Our current corpus [2] consists of 734,488 news articles and 265,512 blog articles, in total around 1 million English-language articles, with an average article length of 405 words.

2.2 Fact Extraction

The initial phase of extracting facts proceeds as follows:

Document Ranking. The system first select the words appearing in the document collection with sufficiently high frequency as topic words, and computes standard TF-IDF weights for each word. The topic words are used to induce document representations. Documents under the same topic are ranked according to the TF-IDF weights of the topic words in each document. The user can pick such topics, and by default, the top-k documents for every topic are selected for further processing.

Coreference Resolution. Pronouns such as "she" are ubiquitous in language and thus entity names often are not explicitly repeated when new facts are expressed in a text. To nevertheless interpret such textual data appropriately, it is thus necessary to resolve pronouns, for which we rely on the Stanford CoreNLP system [5].

Open-Domain Knowledge Extraction. Different sentences within an article tend to exhibit a high variance with regard to their degree of relevance and contribution towards the core ideas expressed in the article. While some express key notions, others may serve as mere embellishments or anecdotes. Large entity network graphs with countless insignificant edges can be overwhelming for end users. To address this, our system computes document-specific TextRank importance scores for all sentences within a document. It then considers only those sentences with sufficiently high scores. From these, it extracts fact candidates as subject-predicate-object triples. Rather than just focusing on named entities (e.g., "Billionaire Donald Trump"), as some previous approaches do, our system supports an unbounded range of noun phrase concepts (e.g., "the snow storm on the East Coast") and relationships with explicit relation labels (e.g., "became mayor of"). The latter are extracted from verb phrases as well as from other constructions. For this, we adopt an open information extraction approach, in which the subject, predicate, and object are natural language phrases extracted from the sentence. These often correspond to syntactic subject, predicate, object, respectively.

2.3 Fact Filtering

The filtering algorithm aims at hiding less representative facts in the visualization, seeking to retain only the most salient, confident, and compatible facts. This is achieved by optimizing for a high degree of coherence between facts with high confidence. The joint optimization problem can be solved via integer linear programming, as follows:

$$\max_{\boldsymbol{x},\boldsymbol{y}} \quad \boldsymbol{\alpha}^\intercal \boldsymbol{x} + \boldsymbol{\beta}^\intercal \boldsymbol{y} \tag{1}$$

$$\text{s.t.} \quad \mathbf{1}^\intercal \boldsymbol{y} \leq n_{\max} \tag{2}$$

$$x_k \leq \min\{y_i, y_j\} \tag{3}$$
$$\forall\ i < j, i, j \in \{1, \ldots, M\},$$
$$k = (2M - i)(i - 1)/2 + j - i$$

$$x_k, y_i \in \{0, 1\}\, \forall i \in \{1, \ldots, M\}, k \tag{4}$$

Here, $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$ with $N = (M+1)(M-2)/2+1$. The y_i are indicator variables for facts t_i: If y_i is true, t_i is selected to be retained. x_k represents the compatibility between two facts $t_i, t_j \in T$ ($i, j \leq M$, $i \neq j$), where $T = \{t_1, \ldots, t_M\}$ is a set of fact triples containing M elements. β_i denotes the confidence of a fact, and $n_{\max}$ is the number of representative facts desired by the user. α_k is weighted by similarity scores $sim(t_i, t_j)$ between two facts t_i, t_j, defined as $\alpha_k = sim(t_i, t_j) = \gamma s_k + (1 - \gamma) l_k$. Here, s_k, l_k denote the semantic similarity and literal similarity scores between the facts, respectively. We compute s_k using the *Align, Disambiguate and Walk* algorithm [6], while l_k are computed using the Jaccard index. $\gamma = 0.8$ denotes the relative degree to which the semantic similarity contributes to the overall similarity score, as opposed to the literal similarity. The constraints guarantee that the number of results is not larger than $n_{\max}$. If x_k is true, the two connected facts t_i, t_j should be selected, which entails $y_i = 1$, $y_j = 1$.

2.4 Conceptual Graph Construction

In order to establish a single connected graph that is more consistent, our system provides an interactive user interface, in which expert annotators can merge potential entities and concepts stemming from the fact filtering process, whose labels present equivalent meanings. They can discover obvious features in the lexical structure of entities or concepts, e.g., Billionaire Donald Trump, Donald Trump, Donald John Trump, Trump, etc. all refer to the same person. For NER, they can use the powerful entity linking ability from a search engine for deciding on coreference. To support the annotators, once again the *Align, Disambiguate and Walk* [6] tool is used for semantically similarity computation between concepts for coreference.

After that, on average, there remains not more than 5 subgraphs that can further be connected for different topics. Hence, users were able to add up to three synthetic relations with freely defined labels to connect these subgraphs into a fully connected graph.

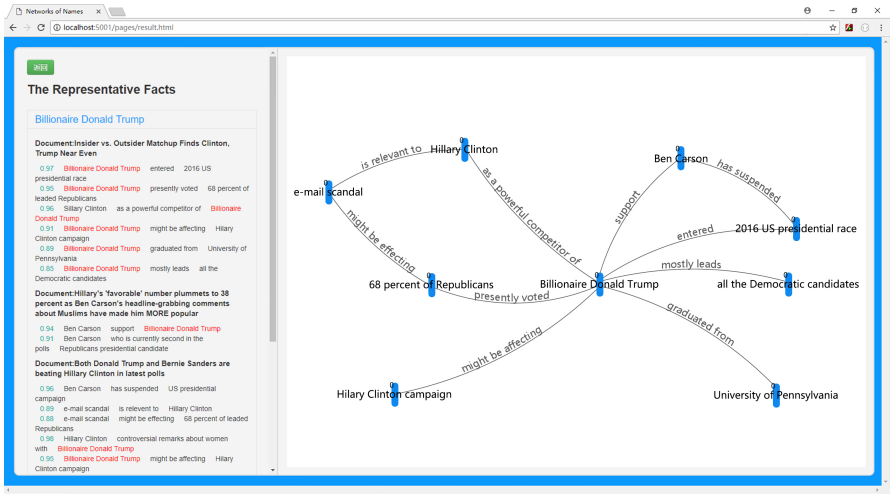

Fig. 1. Example of the user interface: In the left panel, when the user selects the entity "Billionaire Donald Trump" within the set of representative facts extracted from the document topics, the system presents the pertinent entities, concepts, and relations associated with this concept via a graph-based visualization in the right panel, including "Hillary Clinton" as a prominent figure.

The recommended [1] maximum size of a concept graph is 25 concepts, which we use as a constraint. In our evaluation metrics, the coverage rate is the number of topic entities and concepts for which marked as correct divided by the total number of all entities and concepts in the graph. We trained a binary classifier by the topic words with high frequency extracted from different topics to identify the important topic entities and concepts in the set of all potential concepts. We used common features, including frequency, length, language pattern, whether it is named entity, whether it appears in an automatic summarization [4], the ratio of synonyms, with random forests as the model. At inference time for topic concepts, we use the classifier's confidence for a positive classification as the score. We rely on a heuristic to find a full graph that is connected and satisfies the size limit of 25 concepts: We iteratively remove the weakest concepts with relatively lower score until only one connected component of 25 entities and concepts or less remains, which is used as the final conceptual graph. This approach guarantees that the graph is connected with high coverage rate of topic concepts, but might not find the subset of concepts that has the highest total importance score.

3 Related Work

GoWvis [7] is an interactive web application that generates single-document summarizations for a text provided as input, by producing a Graph-of-Words representation. Edges in such graphs, however, merely represent co-occurrences of words rather than specific relationships expressed in the text. The Networks of Names project [3] adopts a similar strategy, but restricted to named entities, i.e., any two named entities co-

occurring in the same sentence are considered related. The Network of the Day project[1] builds on Networks of Names to provide a daily analysis of German news articles. The *news/s/leak* project[2] further extends this line of work by adding access to further corpora and helps journalists to analyse and discover newsworthy stories from large textual datasets. This version also attaches general document keywords as tags to relationships, but does not aim at sentence-level relation semantics as our system.

4 User Interface

Our system is intended to aid users in quickly discerning salient connections in a collection of documents, including via graph-based visualizations. The backend of the system first extracts fact candidates and then attempts to filter out less representative ones, while keeping those deemed most meaningful and important. In the frontend, the system provides compelling visual views, covering multiple steps of the processing pipeline, as shown in Figure 1. A video presenting the user interface is available at `https://shengyp.github.io/vmse`.

5 Acknowledgments

This paper was in part supported by Grants from the Natural Science Foundation of China (No. 61572111), the National High Technology Research and Development Program of China (863 Program) (No. 2015AA015408), a 985 Project of UESTC (No. A1098531023601041) and a Fundamental Research Funds for the Central Universities of China (No. A03017023701). Gerard de Melo's research is funded in part by ARO grant no. W911NF-17-C-0098 (DARPA SocialSim program).

References

1. Cafarella, M.J., Banko, M., Etzioni, O.: Open information extraction from the web (2015)
2. Corney, D., Albakour, D., Martinez, M., Moussa, S.: What do a million news articles look like? In: ECIR. pp. 42–47 (2016)
3. Kochtchi, A., von Landesberger, T., Biemann, C.: Networks of names: Visual exploration and semi-automatic tagging of social networks from newspaper articles. Computer Graphics Forum **33**(3), 211–220 (2014). https://doi.org/10.1111/cgf.12377
4. Li, J., Li, L., Li, T.: Multi-document summarization via submodularity. Kluwer Academic Publishers (2012)
5. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: ACL System Demonstrations (2014)
6. Pilehvar, M.T., Jurgens, D., Navigli, R.: Align, disambiguate and walk: A unified approach for measuring semantic similarity. In: Meeting of the Association for Computational Linguistics (2013)
7. Tixier, A., Skianis, K., Vazirgiannis, M.: Gowvis: A web application for graph-of-words-based text visualization and summarization. In: Acl-2016 System Demonstrations (2016)

[1] http://tagesnetzwerk.de
[2] http://www.newsleak.io/