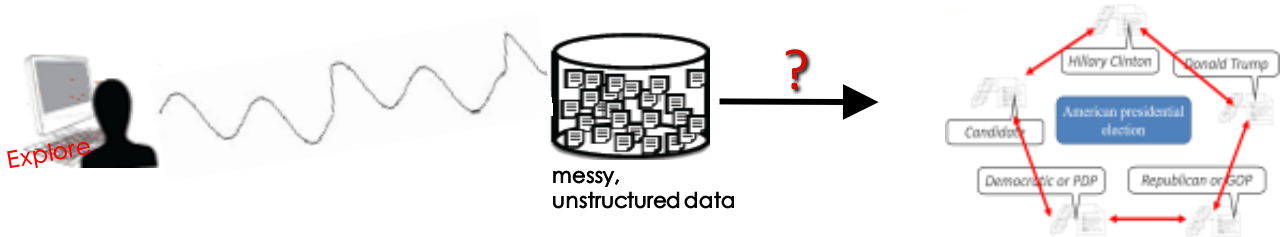
Visualizing Multi-Document Semantics via Open Domain Information Extraction

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

1 University of Electronic Science and Technology of China, China; 2 Shandong University, China; 3 Rutgers University, USA

Motivation

This is challenging to mine and organize meaningful concepts and their semantic connections from a set of related documents under the same topic in the information extraction.



Design and Implementation

Step$_1$

Fact Extraction

- Document Ranking
- Coreference Resolution
- Open-Domain Knowledge Extraction

Step$_2$

Fact Filtering

- Fact filtering algorithm

This is achieved by optimizing for a high degree of coherence between facts with high confidence.

Step$_3$

Conceptual Graph Construction

- Merge potential entities and concepts, whose labels present equivalent meanings.
- **A heuristic algorithm** to find a full graph that is connected and satisfies the size limit of 25 concepts.

Example of the user interface

Backend

- Parsing text documents, pronouns and other forms of coreference are resolved using **Stanford CoreNLP** tool.
- Fact extraction is based on a publicly available system for open information extraction, namely the **KnowItAll Open IE 4.**
- All data is stored in a **MySQL** database.
- Our system is implemented in **Java**, with **Apache Tomcat** as the Web server.

Frontend

- **Topic Keywords Selection:** This view allows users to pick topics based on keywords as queries.
- **Document View:** This view provides a list of documents. Users can browse the full text of the document,
- **Knowledge Extraction View:** Users can extract all facts candidates from multiple documents.
- **Conceptual Graph:** Users can more easily explore the strong connections, and user-selected entities or concepts are highlighted,

For more information make sure to visit the main page: **https://shengyp.github.io/vmse**
Our research group: Statistical Machine Intelligence & LEarning (SMILE Lab), **http://smilelab.uestc.edu.cn**