Interactive visualization of facts extracted from natural language

Anoop Sarkar
Simon Fraser University
Vancouver, Canada
anoop@sfu.ca
Jasneet Sabharwal  Rohit Dholakia  Andrei Vacariu

My collaborators on this project

Ramtin Seraj  Maryam Siahbani  Max Whitney
Summarization
Q. Describe developments in the movement for the independence of Quebec from Canada.

A. Canadian Prime Minister Jean Chretien has ruled out the possibility that the French-speaking province of Quebec could declare independence even if a majority of Quebeckers vote for secession in the next referendum. Premier Lucien Bouchard and the separatist Parti Quebecois have retained control of Canada's largest province, possibly setting the stage for a bitter confrontation with the rest of Canada, and possibly another referendum on independence. Canada's Supreme Court managed to rule on the question of Quebec secession without infuriating leaders on either side of a bitter ideological divide. ...
Summarization by exploiting visualization
The **French Police** questioned **Madame Montpellier**’s **Groundskeeper** and **Housekeeper** at **Madame Montpellier**’s **Home**. Both said a person fitting the description of the bearer of the **Duggan** passport arrived after 6 pm **Aug. 16** to stay the night and that at 11 am **Aug. 17** they found the body of **Madame Montpellier** in her bed with her bedroom door locked from the inside and the bedroom window open. The groundskeeper said the person had arrived in a car with local license plates and that it was gone in the morning. **Madame Montpellier** had apparently been strangled to death.
Topic Models
Language is tough (to visualize)
• the engine.
• the engine caught fire.
• Passengers believed the engine caught fire.
• Passengers reported they saw streaks of flames out of the engine and believed the engine had caught fire.
$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$

$\ 20\ happy\ meal$
Language is tough (to visualize)*

• Words to phrases: very high dimensionality
• Language is compositional and ambiguous
• Abstract concepts
  – So difficult to visualize
• Not pre-attentive
  – Must foveate to read
• Information can be divided into multiple facets

* i247: Information Visualization and Presentation by Marti Hearst
Our approach
Lensing Language

- Semantic parsing of natural language
- Exploit language understanding: *who did what to whom, where, when and how ...*
- Visualize “Things, not Strings”\(^1\): place spatial, temporal and social entities into an intuitive low dimensional space

1: [http://googleblog.blogspot.ca/2012/05/introducing-knowledge-graph-things-not.html](http://googleblog.blogspot.ca/2012/05/introducing-knowledge-graph-things-not.html)
Predicate-centric Ontology

- Semantic Role Labeling to extract predicate argument structures
- Verbose labeling
- Automatically populate a novel predicate-centric ontology
- Using ontology as facet in visualization
  - Easily find uncommon facts
  - Micro-reading in contrast to macro reading
Semantic Role Labeling

• identifying semantic arguments for a verb of a sentence and defining their roles such as who did what to whom, when and where

*The boy* hit *a ball*

*hit*: Predicate

*The boy*: Agent

*a ball*: Patient

\{ Hitter \quad \text{Semantic Roles} \quad \text{Thing hit} \} - Verbose labels
Mary gamely kicked in $5 to John’s bail.

Accuracy: 92%
Proof of concept:
Visualize facts about human history
Crawl Wikipedia for facts

From Wikipedia, the free encyclopedia

Events [edit]

January–December [edit]

- March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam.
- March – The Yorkist King Edward IV returns to England to reclaim his throne.
- April 14 – Battle of Barnet: Edward defeats the Lancastrian army under Warwick, who is killed.

Descriptions

- kill
- found
- defeat

Entity Extraction (NER & hyperlinks)

- Google reverse geo-coding
- Country

Predicates Sense Disambiguation

- roleArg0
- roleArg1

Structured Data

- Locations
- Persons

Event1 (arg0: predicate)

SRL

- event2

Predicate

Structured Data
"arg0": "Emperor Le Thanh Tong",
"arg1": "the Champa Capital",
"event": "capture",
"latitude": 21.03,
"longitude": 105.85,
"country": "Vietnam",
"roleArg0": "getter",
"roleArg1": "thing gotten",
"year": 1471,
"person": "Le Thanh Tong",
"location": {
  "Champa Capital",
  "Vietnam"
},
"description": "March 1 – Emperor Le Thanh Tong captures the Champa Capital, establishing new regions in middle Vietnam."

"Wikipedia_categories": {
  "Vietnamese poets",
  "Lê Dynasty emperors",
  "Southeast Asian countries",
  "15th-century monarchs in Asia",
  ...
}
Domain adaptation

• Map to Wikipedia articles
  – Person
    • Wikipedia categories and infobox: “YEAR births”, “YEAR deaths”, “Kings of*”, “Born”, “Religion”, ...
  – Location
    • Latitude & longitude
  – Organization
    • Wikipedia categories and infobox: “Established in*”, “Companies*”, “Founder”, “Headquarters”, “Employees”, ...

• Ongoing work on machine learning for domain adaptation using Wikipedia Categories
Visualization

2013 CE
April 15 – Two bombs explode at the Boston Marathon, in Boston, Massachusetts in the United States, killing three and injuring 183.

2013 CE
March 27 – Canada becomes the first country to withdraw from the United Nations Convention to Combat Desertification.

2013 CE
February 15 – A meteor explodes over the Russian city of Chelyabinsk, injuring 1,491 people and damaging over 4,300 buildings. It is the most powerful meteor to strike Earth's atmosphere in over a century. The incident, along with a coincidental flyby of an asteroid, prompts international concern regarding the vulnerability of the planet to meteor strikes.
Storyline visualization

Narrative movie chart from [http://xkcd.com/657](http://xkcd.com/657) showing entities arranged in a timeline in clusters based on their interactions
2D maps through embeddings

Events causing death
Comparisons

- theme - 32
- agent - 34
- beginner - 11
- criminal - 8
- patient - 18
Anoop Sarkar, SFU Natural Language Lab, anoop@sfu.ca

- [http://lensingwikipedia.cs.sfu.ca](http://lensingwikipedia.cs.sfu.ca)
- [http://lensingaviation.cs.sfu.ca](http://lensingaviation.cs.sfu.ca)
- [https://github.com/sfu-natlang/lensingwikipedia](https://github.com/sfu-natlang/lensingwikipedia)