

08 – Extracting structured information from text

IA161 Advanced Techniques of Natural Language Processing

V. Kovář

NLP Centre, FI MU, Brno

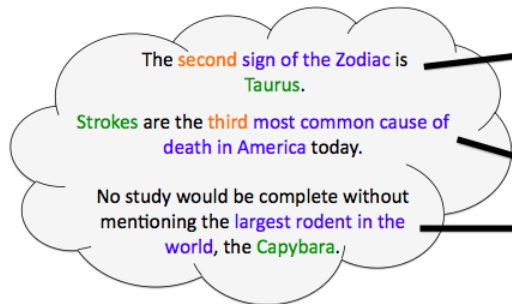
November 11, 2015

- 1 What?
- 2 Why?
- 3 How?
- 4 Notable systems
- 5 Accuracy
- 6 Conclusion

Unstructured Web Text



Structured Sequences



Sign of the Zodiac:

1. Aries
2. Taurus
3. Gemini...

Most Common Cause of Death in America:

1. Heart Disease
2. Cancer
3. Stroke...

Largest rodent in the world:

1. Capybara
2. Beaver
3. Patagonian Cavies

What?

Fed Chairman
Ben Bernanke
said the U.S.
economy...
The euro rose to
\$1.2008,
compared to
\$1.1942
on Tuesday.



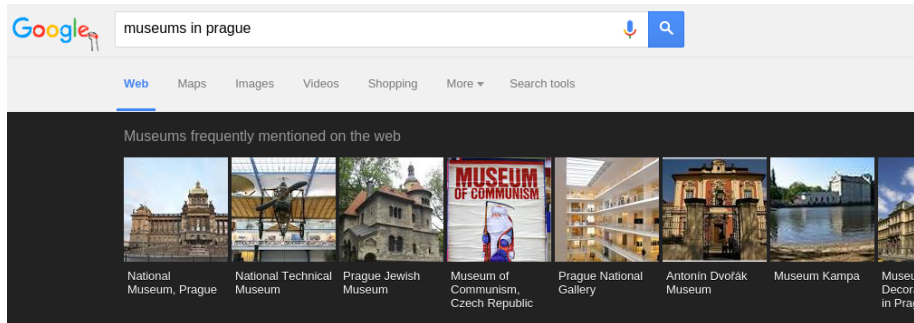
Why?

- Use in subsequent computer applications
 - ▶ information systems
 - ▶ question answering
 - ▶ automatic reasoning
 - ▶ automatic summarization
 - ▶ ...
- Disambiguate and shorten the information
- Highlight particular snippets of information
- *Structured knowledge always better than natural language text*

Successful information extraction systems

(behind the scenes)

- Google



The screenshot shows a Google search interface. The search bar contains the text "museums in prague". Below the search bar, there are navigation tabs for "Web", "Maps", "Images", "Videos", "Shopping", "More", and "Search tools". The "Web" tab is selected. Below the navigation tabs, there is a section titled "Museums frequently mentioned on the web" which displays a horizontal carousel of nine museum images with their respective captions:

- National Museum, Prague
- National Technical Museum
- Prague Jewish Museum
- Museum of Communism, Czech Republic
- Prague National Gallery
- Antonin Dvořák Museum
- Museum Kampa
- Museum Decor in Prag

[Prague Museums - Visitor Information - My Czech Republic](http://www.myczechrepublic.com)

www.myczechrepublic.com > [Prague Guide](#) > [Museums & Galleries](#) ▼

Museums in Prague: National Museum, National Technical Museum and other

- IBM Watson
 - ▶ Jeopardy winner

Big data + “some” intelligence

Successful information extraction systems

- x.ai – automatic personal assistant Amy
 - ▶ agrees automatically on meeting times
 - ▶ recognizes/asks for contact details
 - ▶ operates over Google calendar
- Extracting protein interaction from research texts
- Summarizing and filtering stock market news
- Extracting information about conflicts from news
- Smaller systems for more specialized tasks

How?

- Many variants of the task, depending on application
- Many different systems
- Inconsistent evaluation metrics
- MUC, TREC
 - ▶ Message Understanding Conference + Text REtrieval Conference
 - ▶ series of conferences starting in 80s and 90s
 - ▶ shared tasks + competition among systems
 - ▶ helped standardisation in the field

Usual description of the problem

- Identifying formal relations between objects in text
 - ▶ (has_lecture, Kovář, 11/11/2015 10:00, A219)
 - ▶ (attack, Turkey, ISIS)
- → Division into 2 main sub-tasks
- Named entity recognition
 - ▶ names, companies, time/place expressions
 - ▶ *Turkey, ISIS, A219, Kovář, 11/11/2015*
- Extraction of relations between named entities
 - ▶ also “event extraction” or “filling the gaps”
 - ▶ (who, did, what, when, where)

Methods

- Named entity recognition
 - ▶ finite patterns ('Mr.' capitalized_word+)
 - ▶ list of well-known entities (IBM, Ford, first names)
 - ▶ possibly trained automatically (decision trees, maximum entropy models, hidden Markov models)
- Relation extraction
 - ▶ noun phrase recognition
 - ▶ verb group recognition
 - ▶ → partial parsing
 - ▶ event recognition
 - ▶ rule-based or statistical
 - ▶ also statistical “bag of words” methods

Problems

- Anaphora/coreference resolution
 - ▶ relations may be based on more sentences
 - ▶ “Turkey made a ... It attacked the ISIS in the morning.”
 - ▶ USA, United States, U.S., States
- Many possible expressions for one relation
 - ▶ “Turkey attacked ISIS positions”
 - ▶ “ISIS was attacked by Turkey”
 - ▶ “Turkey joined the war with ISIS”
 - ▶ ...
- Often need for inference
 - ▶ “Thomas J. Watson resigned as president of IBM, and Harriet Smith succeeded him.”

Other notable systems

- Open Information Extraction (OIE), or TextRunner
 - ▶ openie.allenai.org
 - ▶ 100 million web pages
 - ▶ 500 million assertions
- GATE – general architecture for text engineering
 - ▶ gate.ac.uk
 - ▶ huge system for language annotation and all levels of automatic processing
 - ▶ contains a customizable information extraction component
- EFa – Extraction of Facts
 - ▶ nlp.fi.muni.cz/projects/set/efa
 - ▶ in NLP centre at FI
 - ▶ analysis of running text
 - ▶ syntactic analysis
 - ▶ phrase detection
 - ▶ semantic classification of phrases

Accuracy

- Still not very consistent evaluation metrics
- General texts
 - ▶ “fill in the gaps” task (as in MUCs): around 60 %
 - ▶ EFa – precision of phrase detection and classification: 70 %
 - ▶ far from reliable and usable analysis
 - ▶ OIE reports over 80 % *precision*
- Specialized systems
 - ▶ simpler task, e.g. only dates, places, ...
 - ▶ e.g. Amy, the automated personal assistant
 - ▶ much better, human level accuracy

Information extraction – summary

- extracting structured information from text
- named entity detection + relation extraction
- successful in very specialized tasks, not very usable in general tasks

References I



Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007).

Open information extraction for the web.

IJCAI, 7:2670–2676.



Chang, C.-H., Kayed, M., Girgis, M. R., and Shaala, K. F. (2006).

A survey of web information extraction systems.

Knowledge and Data Engineering, IEEE Transactions on,
18(10):1411–1428.



Cunningham, H. (2005).

Information Extraction, Automatic.

Encyclopedia of Language and Linguistics, 2nd Edition.

References II



Fader, A., Soderland, S., and Etzioni, O. (2011).

Identifying relations for open information extraction.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.



Mitkov, R. (2005).

The Oxford handbook of computational linguistics.

Oxford University Press.