11 – Extracting structured information from text

IA161 Advanced Techniques of Natural Language Processing

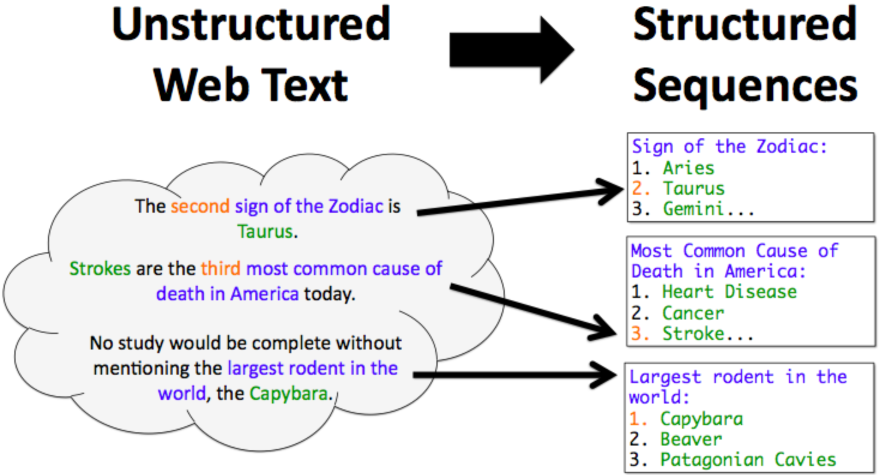
Zuzana Nevěřilová

NLP Centre, FI MU, Brno

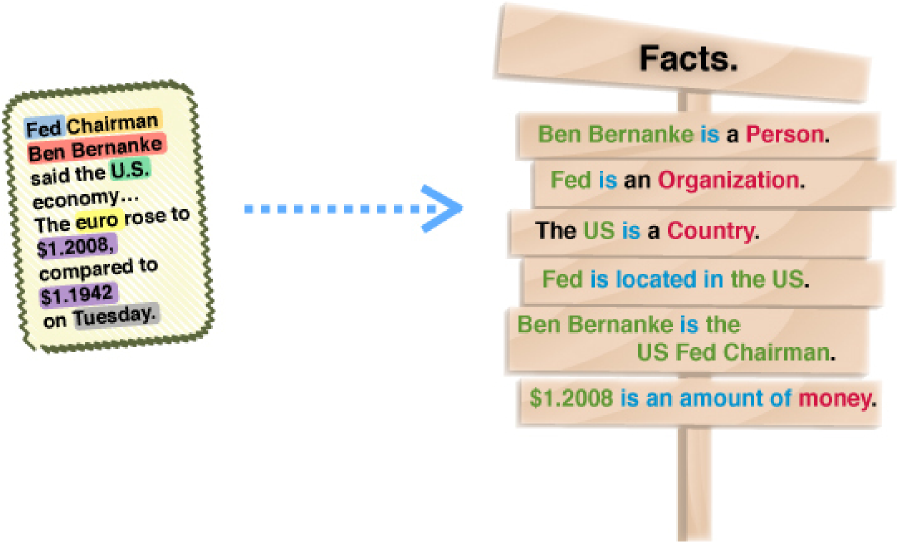
December 17, 2020

1. What?
2. Why?
3. How?
4. Who?

Making Unstructured Information Structured



Information Extraction Goals



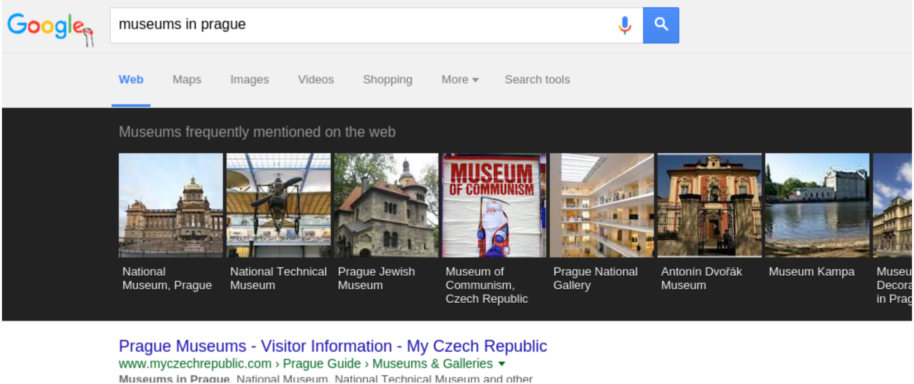
Information Extraction Goals

- Types of Factual Information
  - keywords
  - entities
  - relations
  - events
- Extracted from
  - Different text types: news articles, emails, novels, output from speech recognizer
  - Partially structured resources: lists, databases
  - Different domains or the general domain

Information Extraction Applications

- Direct applications for specific users:
  - financial analysts
  - media analysts
  - PR workers
- Use in subsequent computer applications
  - information systems
  - question answering
  - automatic reasoning
  - automatic summarization
  - ...
- Disambiguate and shorten the information
- Find informational redundancy, aggregate information from several sources

Successful Information Extraction Systems



Google Knowledge Graph (ontologies available at `http://schema.org`)

Successful Information Extraction Systems

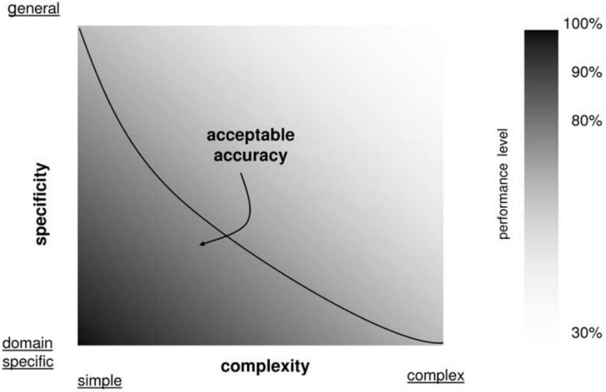
- automatic personal assistants
  - agrees automatically on meeting times
  - recognizes/asks for contact details
  - connects with other applications (e.g. Google Calendar)
- Extracting protein interaction from research texts
- Summarizing and filtering stock market news
- Extracting information about events from news or Twitter
- Smaller systems for more specialized tasks

Information Extraction Evaluation

- Message Understanding Conference + Text REtrieval Conference
- series of conferences starting in 80s and 90s
- shared tasks + competition among systems
- helped standardization in the field
- datasets available
- more recently, many datasets appeared on Kaggle, Zindi, and similar platforms

Information Extraction Approaches

- Specific domain / Complex information
  - precise, narrow requests from small homogeneous corpora
  - weighting/ordering/refining results
- General domain / Simple snippets of information
  - vague request from huge data
  - aggregation of the response



Information Extraction Components

named entity recognition (NE)	finds and classifies names, places, dates, keywords etc.	rocket, Tuesday, Dr Head, Dr Big Head, We Build Rockets Inc.
coreference resolution (CO)	finds identity relations between entities	It = rocket, Dr Head = Dr Big Head
relation extraction (RE)	add description to entities, finds relation between entities (based on CO)	rocket = red shiny, rocket – brainchild – Dr Head, Dr Head – works for – We Build Rockets Inc.
event extraction (EE)	fits RE into event scenarios	rocket launching event

The shiny red rocket was fired on Tuesday. It is the brainchild of Dr Big Head. Dr Head is a staff scientist at We Build Rockets Inc.

Information Extraction Components

named entity recognition (NE)	discussed in detail in lecture 12	Z. Nevěřilová, 17/12/2020, 98071936094, IA161
coreference resolution (CO)	discussed in lecture 04	it = IA161
relation extraction (RE)	discussed in lecture 10 and later in this lecture	Zoom – connect to – 98071936094, IA161 – being taught – 17/12/2020
event extraction (EE)	event recognition, "filling the gaps"	course: name [IA161], date [17/12/2020], zoom ID [98071936094], teacher [Z. Nevěřilová]

domain dependent
tied to scenarios of interest
(ontologies can be used)

The course IA161 takes place every Thursday via Zoom, meeting number 98071936094. The 17th December 2020, it is taught by Zuzana Nevěřilová.

Relation Extraction

In the 10 lecture, a large scale pattern recognition was presented. Here, we focus on processing every single piece of information. The methods overlap heavily, however, the scale is different. So are the criteria for precision/recall.

- noun/verb/adjective/adverbial phrase recognition
- partial parsing
- semantic role labeling (SRL)
- event recognition: actors = noun phrases, action = verb, place = adverbial phrase, time = adverbial phrase
- anaphora and co-reference resolution
- rule-based, statistical, machine learning (Markov models), deep learning (neural networks for sequence modeling)

within a given task, the set of relations is *fixed*
Best MUC results from rule-based or statistical methods: ≈ 75–80% (humans ≈ 90%)

Relation Extraction: Algorithm Example

DIPRE – Sergey Brin's (Dual Iterative Pattern Relation Extraction) [Brin, 1998]

1. initial seed: search for entities connected by well known relations, e.g. `authorship`
2. find occurrences of these pairs over the Internet
3. identify generalized patterns of the contexts of the pairs
4. search for these patterns to discover further names entities with their relationship
5. repeat steps 2 to 4 until no new entities are added

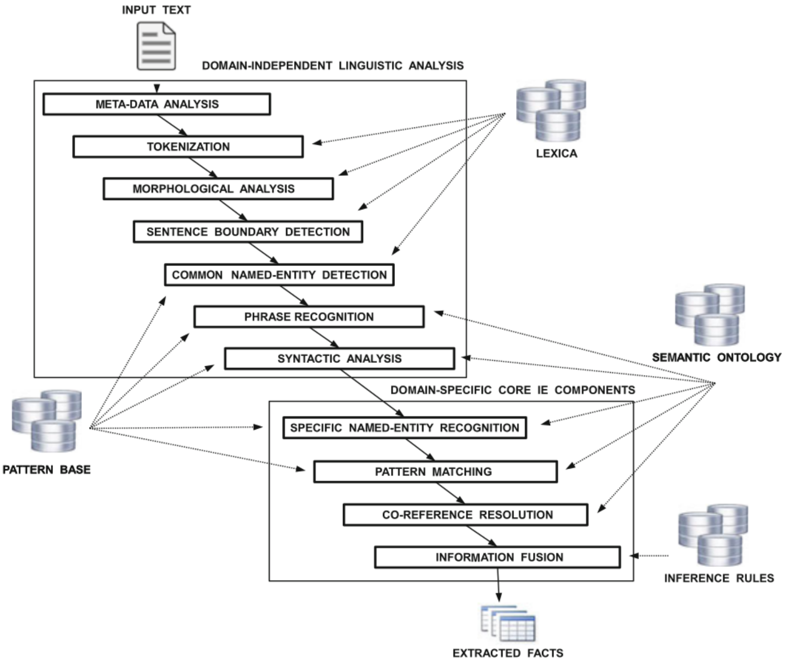
discovering "repeating patterns":
The Godfather was written by Mario Puzo.
Mario Puzo, the author of The Godfather, . . .

Scenario Templates

prototypical outputs

- precision–recall trade-off
- other evaluation metric: slot error rate
  $S = \frac{\mathit{incorrect}+\mathit{missing}}{\mathit{key}}$,
  where *incorrect* is the number of incorrectly assigned slots,
  missing is the number of missing slots,
  and *key* is the total number of slots.

Best MUC results: $\approx$ 60% (humans $\approx$ 80%)


INPUT TEXT
DOMAIN-INDEPENDENT LINGUISTIC ANALYSIS
META-DATA ANALYSIS
TOKENIZATION
MORPHOLOGICAL ANALYSIS
SENTENCE BOUNDARY DETECTION
COMMON NAMED-ENTITY DETECTION
PHRASE RECOGNITION
SYNTACTIC ANALYSIS
LEXICA
SEMANTIC ONTOLOGY
PATTERN BASE
DOMAIN-SPECIFIC CORE IE COMPONENTS
SPECIFIC NAMED-ENTITY RECOGNITION
PATTERN MATCHING
CO-REFERENCE RESOLUTION
INFORMATION FUSION
INFERENCE RULES
EXTRACTED FACTS

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, ...
  - good results in restricted domain (e.g. medical domain where best results are around 86% on i2b2) [Patrick and Li, 2010]
  - much better, human level accuracy

Information extraction: Summary

- extracting structured information from text
- named entity detection + coreference resolution + relation extraction
- event recognition = domain specific, task specific
- successful in very specialized tasks, not very usable in general tasks

Trends:

- social media
- cross-lingual extraction
- open (general) domain

Information Extraction Systems

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

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.

Brin, S. (1998).
Extracting patterns and relations from the world wide web.
In *In WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, pages 172–183.

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.

References II

Cunningham, H. (2005).
Information Extraction, Automatic.
Encyclopedia of Language and Linguistics, 2nd Edition.

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.

References III

Patrick, J. and Li, M. (2010).
High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge.
Journal of the American Medical Informatics Association, 17(5):524–527.

Piskorski, J. and Yangarber, R. (2013).
Information Extraction: Past, Present and Future, pages 23–49.
Springer Berlin Heidelberg, Berlin, Heidelberg.