11 – Extracting structured information from text

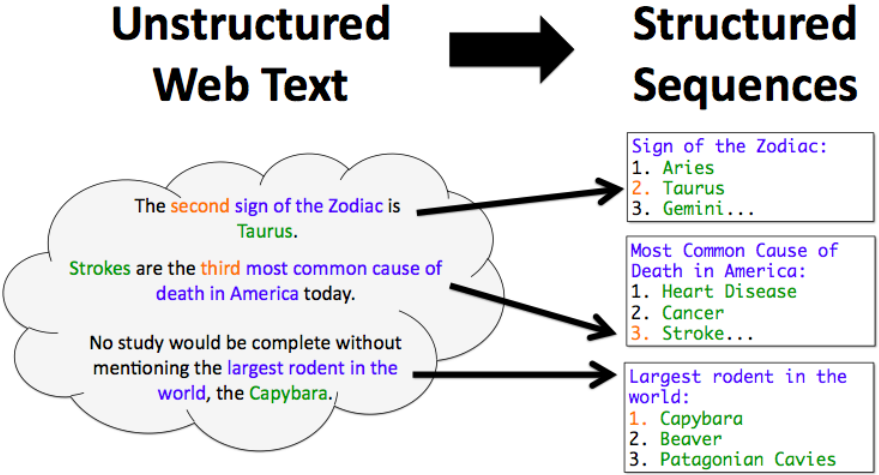
IA161 Advanced Techniques of Natural Language Processing

Zuzana Nevěřilová

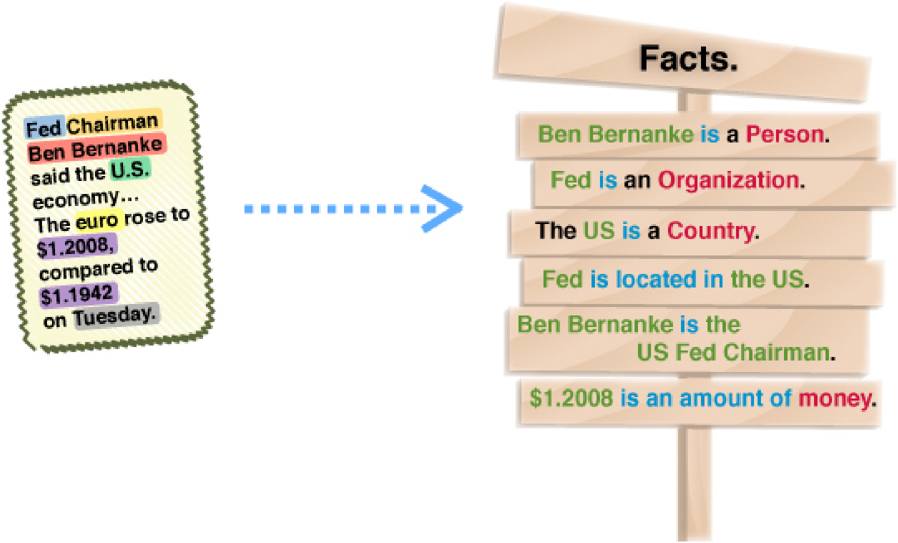
NLP Centre, FI MU, Brno

December 1, 2021

Making Unstructured Information Structured



Information Extraction Goals



Information Extraction Goals: What is a fact

A fact is a statement about important things:

- keywords
- named entities
- date/time
- numbers
- events
- ...

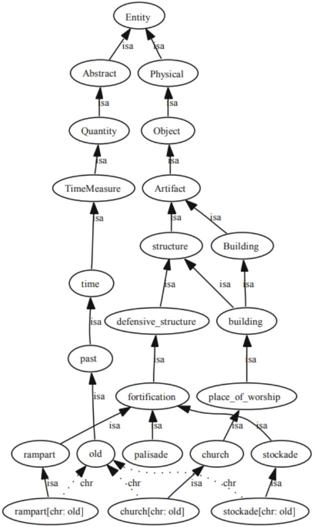
Ontologies

"An ontology is a formal, explicit specification of a shared conceptualization." [Gruber, 1993]

"ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many, or all domains of discourse." [Wikipedia contributors, 2021]

Ontology is a machine readable knowledge base. It usually comes in form of triples
subject – predicate – object

Ontology Example: SUMO/MILO



[Andreasen et al., 2005]

Ontology Example: Schema.org

Taxonomy: Thing > Product > Vehicle > Car

human readable definition: A car is a wheeled, self-powered motor vehicle used for transportation.

Properties: acrissCode, roofLoad
Properties inherited from Vehicle: accelerationTime, cargoVolume, fuelType, numberOfDoors, ...
Properties inherited from Product: brand, color, countryOfOrigin, ...
Properties inherited from Thing: identifier, name, url, ...

Ontology Standards

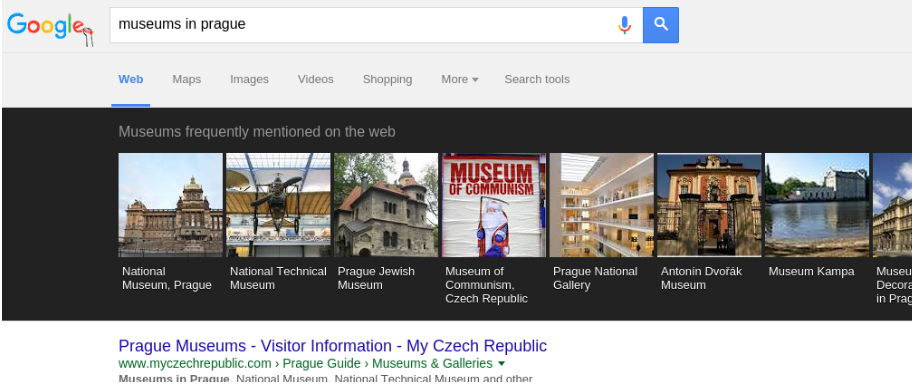
Semantic Web Technologies (W3C standards)

- standard for storing statements: RDF, RDFS
- standard for storing relation types: OWL
- inference based on relation types:
  - Each class if `rdf:subClassOf` of itself.
  - For P being a `owl:TransitiveProperty`, APB and BPC implies APC.
  - ...
- standard query language: SPARQL
- non-standard custom inference languages
- storage in graph databases, relational databases or native triple stores

Information Extraction Applications

- Direct applications for analytical readers:
  - financial analysts
  - media analysts
  - lawyers
  - PR workers
  - students
- Use in subsequent computer applications
  - information systems
  - question answering
  - automatic reasoning
  - automatic summarization
  - dialogue systems
  - ontology engineering
- 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
- IE from social media (noisy)
- Automatic compliance checking with IE from regulatory documents
- Medication IE from clinical notes (dictated)

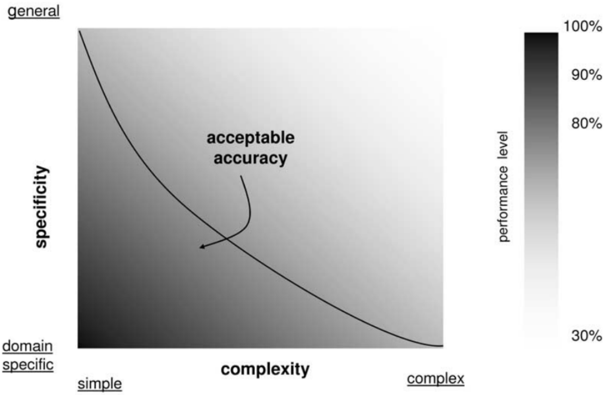
Information Extraction Evaluation

- Message Understanding Conference + Text REtrieval Conference
- SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals
- 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

Dataset overview: `https://github.com/davidsbatista/Annotated-Semantic-Relationships-Datasets`

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 – brain-child – 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 04	Z. Nevěřilová, 01/12/2021, A219, IA161
coreference resolution (CO)	discussed in lecture 12	it = IA161
relation extraction (RE)	discussed in lecture 10 and later in this lecture	IA161 – takes place – A219, IA161 – being taught – 01/12/2021
ontology engineering	what relations are known and expected	Courses take place. Courses are taught by teachers. Teachers are humans.

The course IA161 takes place every Wednesday in room A219.
The 1st December 2021, it is taught by Zuzana Nevěřilová.

Relation Extraction

Scope

- sentence-level
- document-level

Approaches

- hand-crafted rules + statistics
- pattern extraction / bootstrapping (DIPRE, Basilisk [])
- scenarios + filling the gaps
- neural approaches
- machine learning with distant supervision

Best MUC results from rule-based or statistical methods: $\approx$ 75–80% (humans $\approx$ 90%)

Relation Extraction: pattern extraction algorithm

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, ...

Relation Extraction: Basilisk [Thelen and Riloff, 2002]

Generate all extraction patterns in the corpus and record their extractions.
`lexicon = {seed words}`
`i := 0`

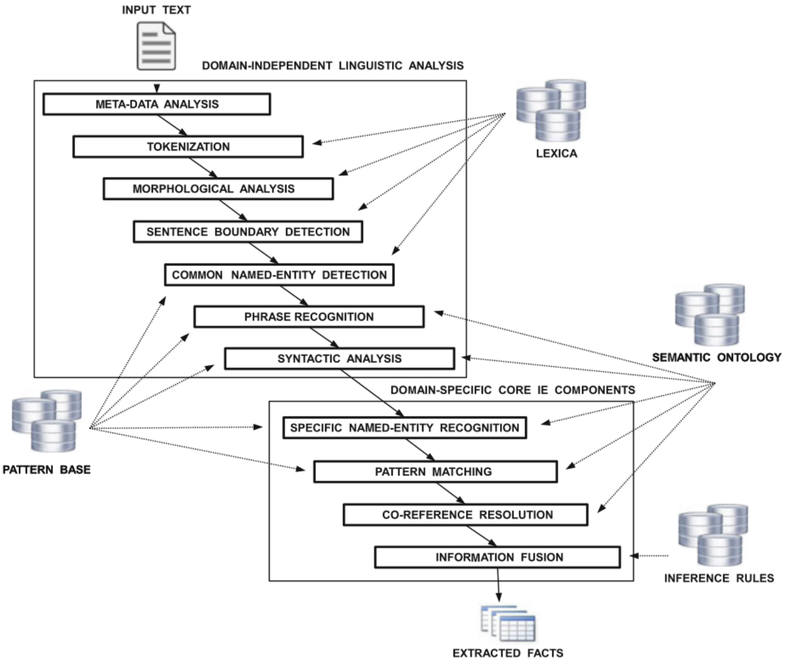
1. Score all extraction patterns
2. `pattern pool` = top ranked 20+i patterns
3. `candidate word pool` = extractions of patterns in `pattern pool`
4. Score candidate words in `candidate word pool`
5. Add top 5 candidate words to `lexicon`
6. `i := i + 1`
7. Go to Step 1.

Scenario Templates

prototypical outputs

- precision–recall trade-off
- other evaluation metric: slot error rate
  $S = \frac{incorrect + missing}{key}$,
  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

Neural Approaches

- convolutional neural networks (CNN): relation classification
- (Bi)LSTM models: shortest dependency[1] path between entities
- Attention mechanism: replace syntactic dependencies
- hierarchical tagging: entity recognition + relation recognition is replaced by entity + relation extraction in one model

Pre-trained models: BERT + graph/hierarchy features

[1] syntactic dependency

Machine Learning with Distant Supervision

In ML, getting the training data is difficult.
Distant Supervision is a database-based approach to collect *positive* examples.

Example

Example
Database knowledge: Barack Obama – married to – Michelle Obama
Mark all sentences with Barack Obama and Michelle Obama as describing the marriage relation.

Problem: negative examples
Possible solution: random samples (e.g. take every sentence mentioning two people as negative marriage relation example.)

Distant supervision = noisy but cheap

Distant Supervision: Removing Noise from Dataset

"If two entities participate in a relation, any sentence that contains those two entities might express that relation." (Mintz, 2009)

- Most of entity pairs have only small number of sentences.
- Lots of entity pairs have repetitive sentences.

[Qin et al., 2018] propose to move false positive examples to negative examples:

- sentence-level FP indicator
- reinforcement learning: classifier training + validation, reward for removing false positives

Accuracy

- 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*
  - best CNN on SemEval2010: 88 %
  - best RNN on SemEval2010: 86.3 %
  - best BERT-based on SemEval2010: 90.2 % [Aydar et al., 2020]
- 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], supervised ML + rule-based approach
  - human level accuracy

Information extraction: Summary

- extracting structured information from text
- named entity recognition + coreference resolution + relation extraction
- event recognition = domain specific, task specific
- successful in very specialized tasks, more difficult in general tasks

Trends:

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

Information Extraction Systems

- Open Information Extraction (OIE)
  - `http://openie.allenai.org`
  - 100 million web pages
  - CALMIE (conjunctions), BONIE (numeric), RelNoun, SRLIE (semantic role labeling)
- 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

Andreasen, T., Bulskov, H., and Knappe, R. (2005).
On automatic modeling and use of domain-specific ontologies.
In Hacid, M.-S., Murray, N. V., Raś, Z. W., and Tsumoto, S., editors, *Foundations of Intelligent Systems*, pages 74–82, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aydar, M., Bozal, O., and Ozbay, F. (2020).
Neural relation extraction: a survey.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007).
Open information extraction for the web.
IJCAI, 7:2670–2676.

References II

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.

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

References III

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.

Gruber, T. R. (1993).
A translation approach to portable ontology specifications.
Knowledge Acquisition, 5(2):199–220.

Mitkov, R. (2005).
The Oxford handbook of computational linguistics.
Oxford University Press.

References IV

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.

Qin, P., Xu, W., and Wang, W. Y. (2018).
Robust distant supervision relation extraction via deep reinforcement learning.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147, Melbourne, Australia. Association for Computational Linguistics.

References V

Thelen, M. and Riloff, E. (2002).
A bootstrapping method for learning semantic lexicons using extraction pattern contexts.
In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 214–221. Association for Computational Linguistics.

Wikipedia contributors (2021).
Ontology (information science) — Wikipedia, the free encyclopedia.
[Online; accessed 28-November-2021].