03 – Extracting structured information from text

IA161 Natural Language Processing in Practice

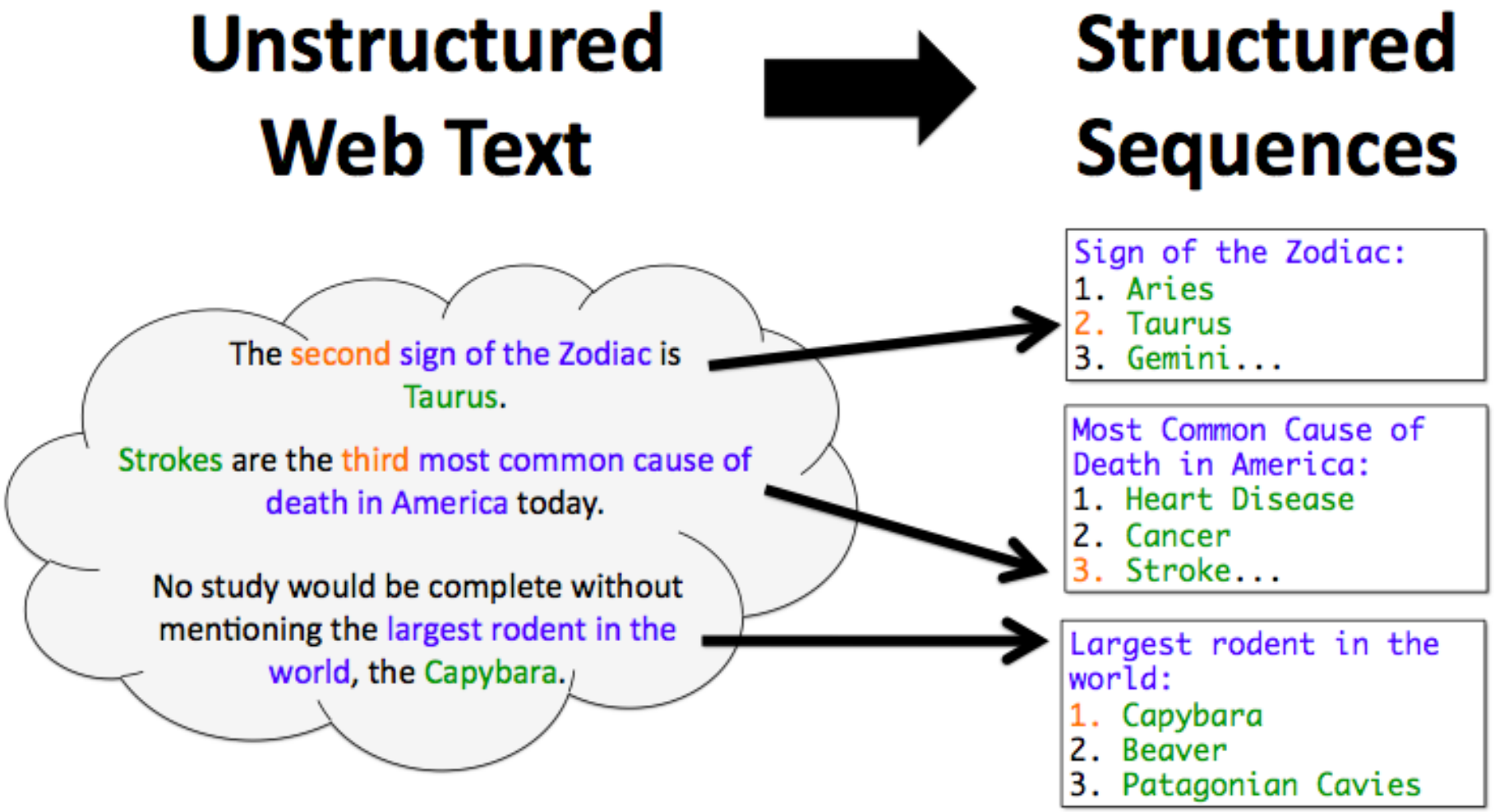
Zuzana Nevěřilová

NLP Centre, FI MU, Brno

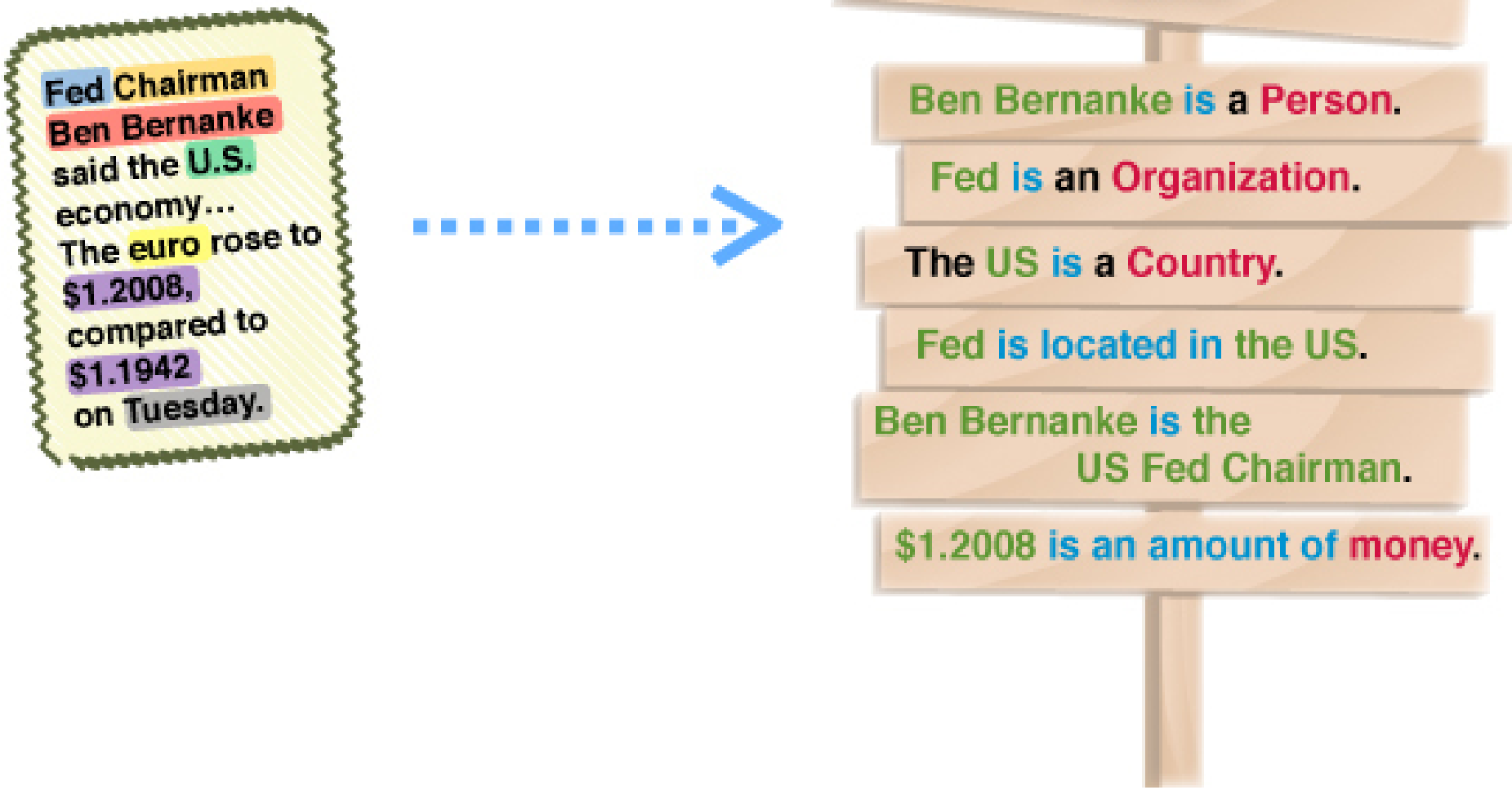
October 3, 2023

1. Information Extraction Goals
2. Knowledge bases and Ontologies
   - Example Ontologies
3. Applications and Example Projects
4. Entity and Relation Extraction Approaches
   - Pattern-based Approach
   - Find Synonyms
   - Find Hypernyms
   - Bootstrapping
   - Machine Learning with Distant Supervision
   - Neural Approaches
5. Evaluation Methods

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
- ...

A statement has a truth value (possibly variable in time and space).
If true, then it is a fact.
A collection of facts someone "understands" is knowledge.
Knowledge is vital for decision-making.
Knowledge can be stored in a knowledge base.

Knowledge bases and Ontologies

"An ontology is a formal, explicit specification of a shared conceptualization."

"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."

Ontologies describe one or more domains in a machine-readable way.
Knowledge bases contain facts about specific individuals.
Apart from the domain knowledge, ontologies can contain facts (as knowledge bases do).

Ontologies and Knowledge Graphs

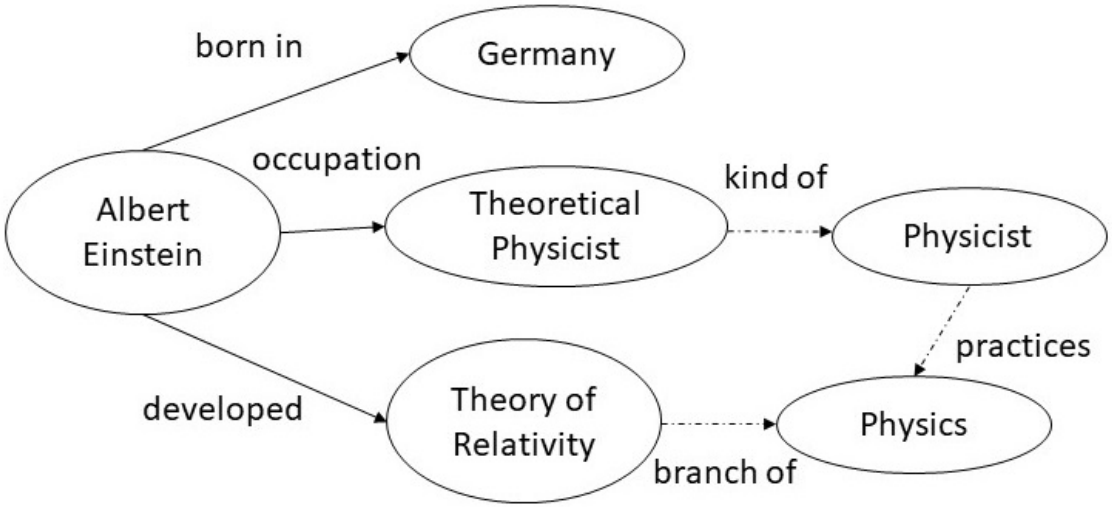
Ontologies are often represented in the form of triples

subject – predicate – object

Ontologies in the Semantic Web:

- subjects, predicates, and objects have unique identifiers (URI)
- the identifiers of subjects and objects are represented as nodes
- the identifiers of predicates are represented as edges
- $\rightarrow$ oriented graph with named nodes and named edges

Example knowledge graph



The ontology collects facts using predefined classes that form a hierarchy.
Physicist → Scientist → Human → LivingBeing
Class instances are connected with other instances with predefined types of edges – properties.
AlbertEinstein → Germany

Ontology Design

The ontology collects facts using predefined classes that form a hierarchy.
Class instances are connected with other instances with predefined types of edges – properties.
Concepts differ from their labels ("Germany" and "Deutschland" are labels of the concept Germany).

Ontology Design

Ontologies can contain statements about classes and statements about class members (instances).
Statements about classes:

- Physicists are Scientists
- Scientists are Humans
- Humans have Place of Birth
- Place of Birth is a Location

Statements about instances:

- Albert Einstein was a Physicist
- Albert Einstein was born in Germany

Possible inferences:

- Albert Einstein was a Scientist
- Albert Einstein was a Human
- Germany is a Location

Ontology Design

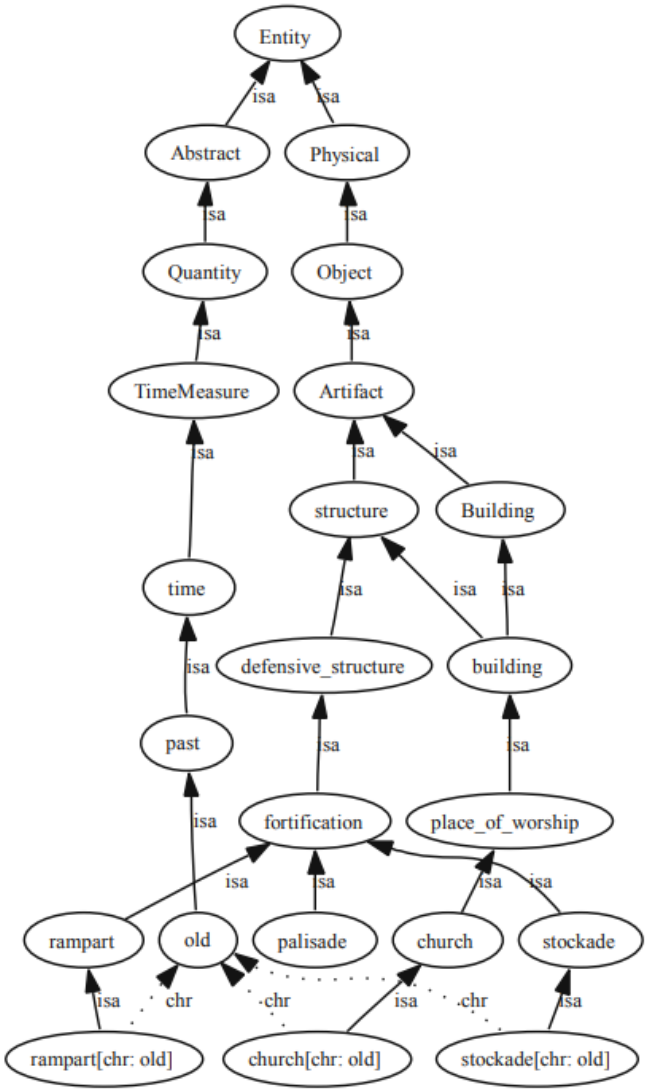
Ontologies can contain statements about classes and statements about class members (instances).
Statements about classes:

- Physicists are Scientists
- Scientists are Humans
- Humans have Place of Birth
- Place of Birth is a Location

. . . formal, explicit specification of a shared conceptualization . . .

Statements about classes are universal, so they are often shared across ontologies.

Ontology Example: SUMO/MILO



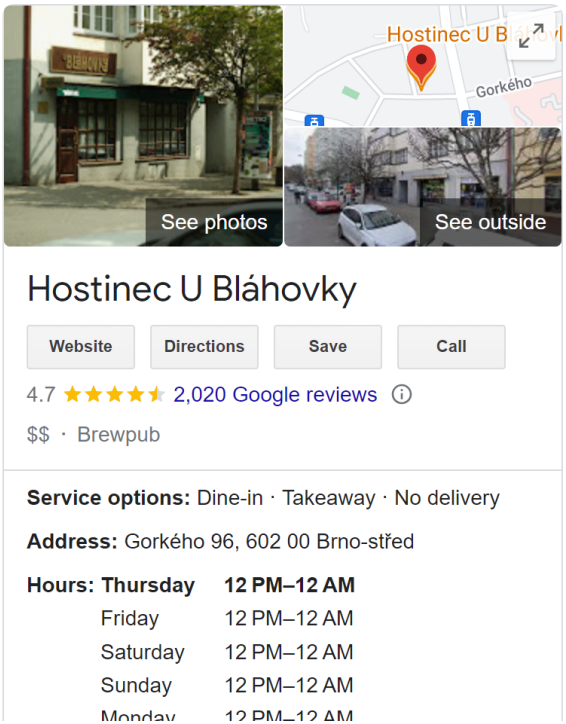
[Andreasen et al., 2005]

Ontology Example: Schema.org

Taxonomy: Thing > Place > LocalBusiness

human-readable definition: A particular physical business or branch of an organization.

Properties: openingHours, paymentAccepted
Properties inherited from Place: address, geo, photo ...
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 (impact on inference)
  - Each class if `rdf:subClassOf` of itself.
  - For P being a `owl:TransitiveProperty`, APB and BPC implies APC.
  - ...
- standard query language: SPARQL

Ontologies are stored in graph databases, relational databases or native triple stores.

Information Extraction Applications

- Direct applications for analytical readers:
  - financial/media analysts
  - lawyers
  - biologists, bio-medics
- Use in subsequent computer applications
  - query expansion (dog $\rightarrow$ poodle, terrier)
  - form value extraction
  - automatic summarization
  - information aggregation
  - question answering, dialogue systems
  - automatic reasoning, support for decision making

Information Extraction Systems

- Open Information Extraction (OIE)
  - `https://demo.allennlp.org/open-information-extraction`
  - 100 million web pages
  - CALMIE (conjunctions), BONIE (numeric), RelNoun, SRLIE (semantic role labeling)
- IBM Grapher
  `https://github.com/IBM/Grapher` [Melnyk et al., 2022]
- GPT-based knowledge graph generation
  - GraphGPT (`https://graphgpt.vercel.app/`)
  - SchemaGPT (`https://github.com/u-hubar/SchemaGPT`)
  - try your own prompt, e.g.,
    "Generate an RDF rendition of the following using Turtle Notation, terms from Schema.org, and hash-based relative hyperlinks for subject and object denotation. Use a single code-block for response output."[1]

[1] `https://medium.com/openlink-software-blog/using-chatgpt-to-generate-a-course-description-knowledge-graph-using-rdf-d`

Entity and Relation Extraction Approaches

Entity extraction approaches

- named entity recognition (NER)
- "all" noun phrases (coreference, idioms)

Relation extraction approaches

- hand-crafted rules (syntactic patterns) + statistics
- pattern extraction + bootstrapping (DIPRE, Basilisk [Thelen and Riloff, 2002])
- machine learning with distant supervision
- neural approaches (two-step + one-step)

Pattern Recognition

regular expression matches part-of-speech (we need a POS tagger)

Example

NP {,} especially {NP, }* {or |and} NP
...most *European countries*, especially *France*, *England*, and *Spain*.
France is a European country
England is a European country
Spain is a European country

Example

e.g. {NP,}* {and |or} NP.
. . . e.g., apples, bananas, or pears.
related terms

Pattern Recognition

Example

NP such as {NP, }* {and |or} NP

common *domestic animals* such as the *ferret* and the *fancy rat*

ferret is a domestic animal

(fancy) rat is a domestic animal

in areas with a long history of *mining* such as *South-west England*

South-west England is a area

South-west England is area of mining

- POS-tag words
- remove stopwords
- detect optional adjunct phrases
- detect named entities

Pattern Recognition

No.	Pattern	Number of occurrences	Number of relevant occurrences	Intermediary precision (%)
1.	other than	168	164	97.6
2.	especially	120	90	75
3.	principally	11	6	54.5
4.	usually	18	14	77.8
5.	such as	2470	1950	78.9
6.	in particular	78	48	61.5
7.	e(.)g(.)	280	216	77.1
8.	become	780	510	66.7
9.	another	92	72	78.3
10.	notably	76	42	55.3
11.	particularly	130	80	61.5
12.	except	13	4	30.8
13.	called	270	220	81.5
14.	like	1600	1300	81.3
15.	including	670	430	64.2

https://doi.org/10.1007/978-3-540-87391-4_7

Corpus Query

- special case of pattern recognition, CQL query
- bigger data at hand, highly variable texts

Example

je/jsou

```
2:[k="k1"&c="c1"] ([lc=","] [k="k1"])*
([lc="a"|lc="i"|lc="nebo"|lc="či"] [k="k1"])?
[lemma_lc="být"&tag="k5eAaImIp3.*"&lc!="ne.*"]
([k="k1"&c="c[1246]"] [k="k2"]{0,2})?
1:[k="k1"&c="c[1246]"]
```

experiment on domain dictionary: precision 40 %, when limited to dictionary terms 52 %

Multiple Translation

using translation equivalents from a multilingual dictionary to provide synonyms

Example

stůl = table
table = stůl, stolek
stůl = stolek

Synonym Transitivity

- expanding relations based on existing relations (transitive closure)

Example

city = town, town = municipality
$\Rightarrow$ city = municipality

Hypernym Extraction

apple, pear, orange $\rightarrow$ fruit
Crucial in terminology dictionaries and ontologies (hierarchy, inference)

- Used in definitions (e.g., apple is a fruit that . . .) – [Wang and Hirst, 2012]
- Pattern-based hypernym extraction: [Grefenstette, 2015, Lefever et al., 2014, Sang and Hofmann, 2009]
- Pattern-based + distributional approach: [Schropp et al., 2013]

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.

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 entity pairs have only a 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

Neural Approaches

Two step procedure:
Step 1: term extraction
Step 2:

- Convolutional neural networks (CNN): relation classification [Li et al., 2020]
- (Bi)LSTM models: shortest dependency[2] path between entities
- Attention mechanism: replace syntactic dependencies [Shen and Huang, 2016]

One step procedure:
Hierarchical tagging: entity recognition + relation recognition is replaced by entity + relation extraction in one model

[2] syntactic dependency

Information Extraction Evaluation Datasets

- series of conferences starting in the 80s and 90s: Message Understanding Conference (MUC) + Text REtrieval Conference (TREC) with shared tasks + datasets available
- SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals
- SemEval-2015 Task 17: Taxonomy Extraction Evaluation (TExEval)
- TOEFL test evaluation
- WebNLG `https://huggingface.co/datasets/web_nlg`
- Wikidata verbalisation dataset (WDV, `https://doi.org/10.6084/m9.figshare.17159045.v1`) [Amaral et al., 2022]
- TEKGEN `https://github.com/google-research-datasets/KELM-corpus`
- `https://paperswithcode.com/task/relation-extraction`

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

TOEFL test evaluation

- evaluation by solving TOEFL synonym test
- Choose synonym for *fabricate*.
  - construct, alter, select, demonstrate
- build synonym set for each word
- detect overlap
- success rate 88 %

Accuracy

- General texts
  - "fill in the gaps" task (as in MUCs): around 60 %
  - 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 a 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
  structure information + meaning = knowledge
- successful in very specialized tasks, more difficult in general tasks

Trends:

- from multiple steps to one-step extraction
- noisy extraction (social media)
- cross-lingual extraction
- open (general) domain

References I

Amaral, G., Rodrigues, O., and Simperl, E. (2022).
WDV: A broad data verbalisation dataset built from wikidata.
In Sattler, U., Hogan, A., Keet, M., Presutti, V., Almeida, J. P. A., Takeda, H., Monnin, P., Pirrò, G., and d'Amato, C., editors, *The Semantic Web – ISWC 2022*, pages 556–574, Cham. Springer International Publishing.

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.

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.

Grefenstette, G. (2015).
INRIASAC: Simple Hypernym Extraction Methods.
arXiv preprint arXiv:1502.01271.

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

References III

Lefever, E., Van de Kauter, M., and Hoste, V. (2014).
Evaluation of Automatic Hypernym Extraction from Technical Corpora in English and Dutch.
In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 490–497.

Li, Q., Li, L., Wang, W., Li, Q., and Zhong, J. (2020).
A comprehensive exploration of semantic relation extraction via pre-trained CNNs.
Knowledge-Based Systems, page 105488.

Melnyk, I., Dognin, P., and Das, P. (2022).
Knowledge graph generation from text.
In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*.

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.

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

Sang, E. T. K. and Hofmann, K. (2009).
Lexical patterns or dependency patterns: which is better for hypernym extraction?
In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 174–182. Association for Computational Linguistics.

Schropp, G., Lefever, E., and Hoste, V. (2013).
A Combined Pattern-based and Distributional Approach for Automatic Hypernym Detection in Dutch.
In *RANLP*, pages 593–600.

Shen, Y. and Huang, X.-J. (2016).
Attention-based convolutional neural network for semantic relation extraction.
In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2526–2536.

References VI

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.

Wang, T. and Hirst, G. (2012).
Exploring patterns in dictionary definitions for synonym extraction.
Natural Language Engineering, 18(03):313–342.

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