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

Unstructured Web Text

The second sign of the Zodiac is Taurus.

Strokes are the third most common cause of death in America today.

No study would be complete without mentioning the largest rodent in the world, the Capybara.

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
Information Extraction Goals

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

Facts.

Ben Bernanke is a Person.
Fed is an Organization.
The US is a Country.
Fed is located in the US.
Ben Bernanke is the US Fed Chairman.
$1.2008 is an amount of money.
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

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

*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**

<table>
<thead>
<tr>
<th>Component</th>
<th>Details</th>
<th>Instructor</th>
<th>Date</th>
<th>Zoom ID</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>named entity recognition (NE)</td>
<td>discussed in detail in lecture 12</td>
<td>Z. Nevěřilová</td>
<td>17/12/2020</td>
<td>98071936094, IA161</td>
<td></td>
</tr>
<tr>
<td>coreference resolution (CO)</td>
<td>discussed in lecture 04</td>
<td>it = IA161</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>relation extraction (RE)</td>
<td>discussed in lecture 10 and later in this lecture</td>
<td>Zoom – connect to –</td>
<td>98071936094,</td>
<td>IA161 – being taught – 17/12/2020</td>
<td></td>
</tr>
<tr>
<td>event extraction (EE)</td>
<td>event recognition, “filling the gaps”</td>
<td>course: name [IA161], date [17/12/2020], zoom ID [98071936094], teacher [Z. Nevěřilová]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

↑ domain dependent
   tied to scenarios of interest
   (ontologies can be used)
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: $\approx 75$–$80\%$ (humans $\approx 90\%$)
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, . . .
prototypical outputs

- precision–recall trade-off
- other evaluation metric: slot error rate

\[ S = \frac{\text{incorrect} + \text{missing}}{\text{key}} \]

where \textit{incorrect} is the number of incorrectly assigned slots, \textit{missing} is the number of missing slots, and \textit{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

PATTERN BASE

DOMAIN-SPECIFIC CORE IE COMPONENTS

- SPECIFIC NAMED-ENTITY RECOGNITION
- PATTERN MATCHING
- CO-REFERENCE RESOLUTION
- INFORMATION FUSION

INFERENCERULES

SEMANTIC ONTOLOGY

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](http://openie.allenai.org)
  - 100 million web pages
  - 500 million assertions

- GATE – general architecture for text engineering
  - [http://gate.ac.uk](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
  - in NLP centre at FI
  - analysis of running text
  - syntactic analysis
  - phrase detection
  - semantic classification of phrases
Open information extraction for the web. 
*IJCAI*, 7:2670–2676.

Extracting patterns and relations from the world wide web. 

A survey of web information extraction systems. 
*Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1411–1428.

