INFORMATION EXTRACTION

Jelena Jovanović

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Outline

- What is Information Extraction (IE)?
- Application domains
- Five main types of IE tasks
- Factors affecting the performance of IE systems
- Features and performance of the main IE tasks

What is Information Extraction?

- A branch of the Natural Language Processing (NLP) field
- It integrates methods and techniques for text analysis aimed at extracting structured information from unstructured text
- The information to be extracted is typically related to predefined types of entities, relationships and events

What is Information Extraction?

An example

"Novak Djokovic extended his unbeaten record at the 2012 China Open in Beijing on Tuesday, but needed three sets to defeat qualifier Michael Berrer of Germany.

The Serb progressed to the second round with a 6-1, 6-7 (3/7), 6-2 definitive victory…"

Information that could be extracted:

- Novak Djokovic; 2012 China Open; Beijing; Tuesday; …; the Serb.
- *Novak Djokovic* and *the Serb* refer to the same entity
- unbeaten (record); definitive (victory)
- (Novak Djokovic) qualifier (Michael Berrer)
- qualify for the 2nd round of 2012 China Open

IE vs. Information Retrieval (IR)

EI & IR: two similar, but still very different tasks:

- IR system finds and returns (potentially) relevant *documents* and presents them to the user
  - An example: Web search engines such as Google or Bing
- IE system analyzes documents, and extracts and presents only those *pieces of information* it estimates (potentially) relevant to the user
  - An example: IBM Watson

Application domains

Advanced Web search

"*Google is undergoing a major, long-term overhaul of its search-engine, using what's called semantic Web search…*

*…people familiar with the initiative say that Google users will able to browse through the company's '**knowledge graph**,' or its ever-expanding **database** of information **about 'entities'** – people, places and things – **the 'attributes' of those entities** and **how** different **entities are connected** to one another*."

What Google's Search Changes Might Mean for You

Wall Street Journal, March 14, 2012

Application domains

Advanced Web search

*"At the core of Microsoft's work to create a state-of-the-art Bing digital assistant is **Satori**, a **knowledge repository** of more than a billion objects digested in the past 3.5 years…*

*…Satori catalogs **entities** and the **associated data and relationships** among them...*

*…Satori is a **self-learning system** that is running every day and learning more, adding 28,000 DVDs of content every day…*

…Bing search and Windows already are using Satori's knowledge repository…"

Microsoft's Bing seeks enlightenment with Satori
CNET News, July 30, 2013

Application domains

Advanced search of social networks

"Facebook is building a rich stock of knowledge that could make its software smarter and boost the usefulness of its search engine…

*…Entities such as colleges and employers are **learned from data** typed **in profile pages**; businesses, movies, fictional characters, and other concepts are **learned from fan pages** created by Facebook users. … **analyzing** many **employment histories** on the site allows Facebook's search engine to know that a search for "software engineers" should also return people who say they are "coders."*

Facebook Nudges Users to Catalog the Real World

MIT Technology Review, February 27, 2013

Application domains

Business analytics

- Extraction of information of interests for organizational management and decision making, e.g., information about
  - geo-political and macro-economic events, and/or
  - events of interest for particular companies and/or their brands
- Extracted information is typically used as a feed for a Business Rules Engine
- Examples:
  - RavenPack News Analytic (http://www.ravenpack.com/)

Application domains

Marketing & PR

- Social media monitoring
- Reputation management
- Social customer relationship management
- Examples:
  - Safesforce Marketing Cloud (ex. Radian6; link)
  - Lithium Social Intelligence product (link)

Application domains

Marketing & PR (cont.)

"*It's fine to have a Facebook fan page or count 'Likes' but really, that's looking at a tiny, narrow portion of a social strategy*"

"There are millions of conversations going on related to retailers, their products and their brands. They need to be aware of and participate in that breadth of the social web."

Izvor: http://bit.ly/yg7hTO

Application domains

Online advertising

- Context-aware analysis of the Web page content and extraction of:
  - the type of text,
  - entities (persons, locations, companies, brands,…),
  - sentiment/emotions expressed in the text,
  - the overall message the text aims to communicate
- Thus extracted information serves as an input for the recommendation of ads for the analyzed Web page
- An example: ADmantX (http://www.admantx.com/)

Five main types of IE tasks

- Named Entity recognition
  - recognition of different kinds of entities mentioned in the text (persons, organizations, dates, currencies, …)
- Co-reference resolution – takes one of the following forms:
  - anaphoric resolution
    - Ex. identify that in the text: "*Tom* is my best friend. I know *him* since we were kids." pronoun 'him' refers to the noun 'Tom'
  - proper noun resolution
    - Ex. determine that the following terms refer to the same entity: 'IBM', 'IBM Europe', 'International Business Machines Ltd.', . . .

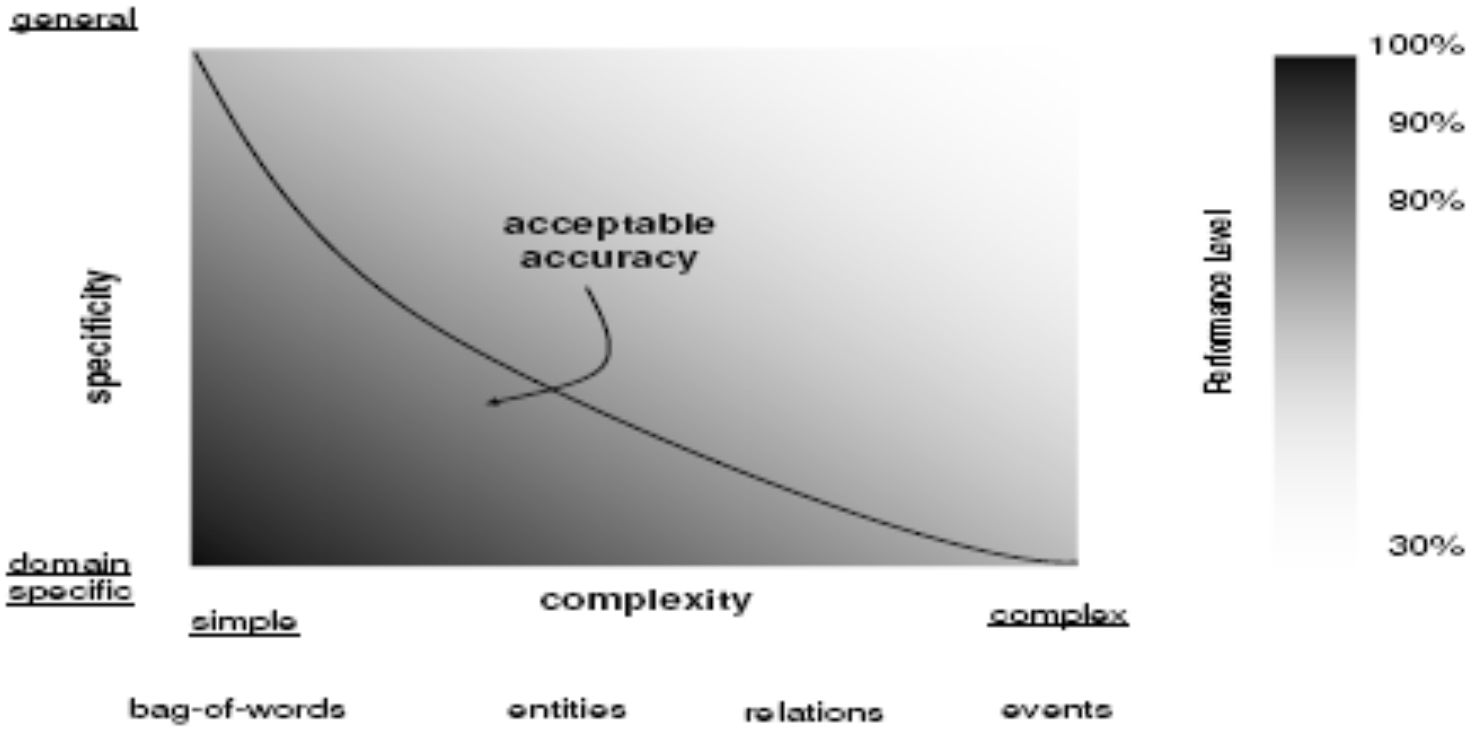
Five main types of IE tasks

- Descriptions resolution
  - What attributes do the entities have?
- Relations resolution
  - What kind of relationships exist among the entities?
- Events resolution
  - Identification of events the entities participate in

IE: Factors affecting the performance

- Features of the specific IE task
  - The type of the text – the kind of text the task is related to
    - for example: Wall Street Journal articles, or email messages, or novels, or the output of a speech recognizer
  - Topic or domain – the broad subject-matter of the text
    - for example: world events, or seminar announcements, or financial news
  - The writing style
    - formal/informal, with jargon or jargon free,…
  - The type of information to be extracted
    - For example, persons, companies, relationship between people and companies,...

IE: Factors affecting the performance



Performance trade-off
related to specificity and complexity of an IE task

Source: H. Cunningham, Information Extraction, Automatic. Encyclopedia of Language and Linguistics, 2nd Edition, Elsevier. 2005.

IE: Evaluation measures

Two most frequently used measures for determining/ presenting the performance of an IE (or IR) system:

- ***Precision*** – Are all the extracted pieces of information relevant?
- ***Recall*** – Have all relevant information pieces being extracted?

	Correct	Incorrect
Extracted	A	B
Not extracted	C	D

Precision = A / (A U B)

Recall = A / (A U C)

IE: Evaluation measures

Increase in precision almost always goes at the expense of recall:

- we can develop systems that do not make many mistakes, but that miss quite a lot of occurrences of relevant information;
- alternatively we can push up recall and miss less, but at the expense of making more mistakes.

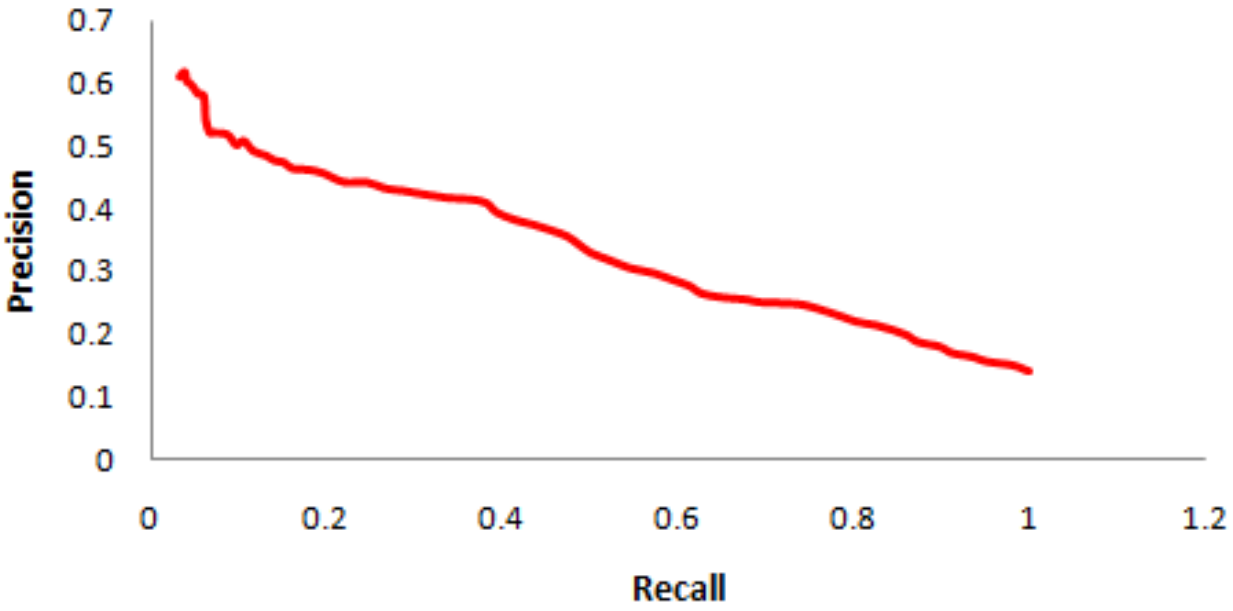


Image source: http://groups.csail.mit.edu/cb/struct2net/webserver/images/prec-v-recall-v2.png

Features and performance of the main IE tasks

Named Entity Recognition

- Operates almost at the human level
  - Up to ~ 95% precision
  - Even people do not perform this task with 100% precision
- Weakly domain dependent
  - E.g., changing the subject-matter of the texts being processed from financial news to other types of news would involve only some slight changes to the system
- Very dependent on the type of the text being analyzed
  - E.g., the Stanford NER drops from 90.8% F1 to 45.88% when applied to a corpus of tweets*

* Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. (2011) 359-367

Co-reference resolution

- Main application area:
  - Assignment of descriptive information dispersed throughuot the text to the entities this information refers to
- Performance:
  - Imprecise task
  - Results tend to greatly vary from domain to domain – highly domain dependent task
  - Depending on the domain, precision can reach the 50-60% level

Description resolution

- Formal name: *Template Element construction*
- Makes use of the results of the two previously described tasks
- Adds descriptive information to the recognized entities
- Performance
  - Precision level is up to 80%
  - Human performance on this type of task is about 95%
- Portability
  - Weakly domain dependent task
  - Highly dependent on the type of the text

Relation resolution

- Formal name: *Template Relation construction*
- Makes use of the results of the previously described tasks
- It is about identification of relationships between named entities
  - Examples:

    relationship between an employee and the company he/she works for,

    family relationships between two or more persons,

    relationships between two companies, …
- Performance
  - Precision level is up to 75%
- Portability
  - Weakly domain dependent task

Events resolution

- Formal name: *Scenario Templates recognition*
- Makes use of the results of the previously described tasks: it connects entities, their descriptions and relationships into descriptions of events
- Performance
  - Precision level is up to ~ 60%
- Portability
  - Domain independent
  - Closely tied to the scenarios of interest for the given user

(Anonymous) questionnaire for your critique, comments, suggestions:

http://goo.gl/cqdp3I