COMP348 — Document Processing and the Semantic Web

Week 6 Lecture 1: Information Extraction

Diego Mollá

Department of Computer Science
Macquarie University

COMP348 2017H1

Programme

1. Information Extraction
2. Named Entity Recognition

Reading

- NLTK Chapter 7. http://nltk.org/book/ch07.html

Some Useful Extra Reading

- Chapter 3 “Searching for Named Entities” of Barrière (2016) “Natural Language Understanding in a Semantic Web Context”
- Chapter 3 “Named Entity Recognition and Classification” of Maynard et al (2016) “Natural Language Processing for the Semantic Web”.
- David Nadeau, Satoshi Sekine (2007). A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes* 30:1.

Programme

1. Information Extraction
2. Named Entity Recognition

The Motivation for Information Extraction

Observations

- Most of the information is contained in text in human languages and not in databases or similar structured formats.
- Most of new information is now stored in digital form.

Conclusion

You're missing out on a lot of good stuff if you can't get answers from all that digital information written in human languages.

The Motivation for Information Extraction

Observations

- Most of the information is contained in text in human languages and not in databases or similar structured formats.
- Most of new information is now stored in digital form.

Conclusion

You're missing out on a lot of good stuff if you can't get answers from all that digital information written in human languages.

What Information Extraction is About

The Problem

Extract well-defined pieces of information (for example named entities or events) from collections of documents.

The Goal

To populate a template or database.

Particularities

- Typically, most of the information in a document is ignored.
- IE can be contrasted with earlier goals of building story understanding systems, where broad and deep coverage is needed.

What Information Extraction is About

The Problem

Extract well-defined pieces of information (for example named entities or events) from collections of documents.

The Goal

To populate a template or database.

Particularities

- Typically, most of the information in a document is ignored.
- IE can be contrasted with earlier goals of building story understanding systems, where broad and deep coverage is needed.

What Information Extraction is About

The Problem

Extract well-defined pieces of information (for example named entities or events) from collections of documents.

The Goal

To populate a template or database.

Particularities

- Typically, most of the information in a document is ignored.
- IE can be contrasted with earlier goals of building story understanding systems, where broad and deep coverage is needed.

An Example Document I

From MUC-4

San Salvador, 19 Apr 89 (ACAN-EFE) – [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

...

Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador.

...

Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle.

...

According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

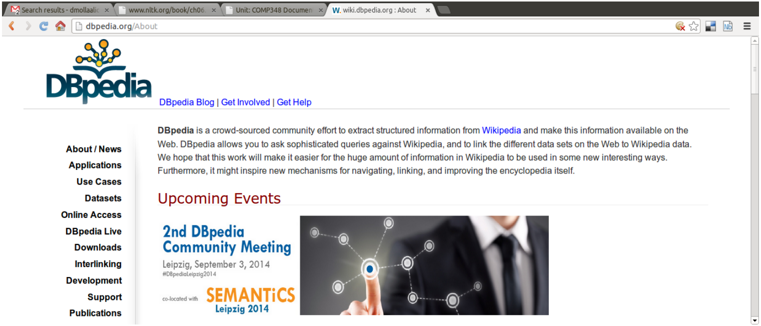
An Example Document II

A Corresponding Filled Template

Incident: Date	19 Apr 89
Incident: Location	El Salvador: San Salvador (CITY)
Incident: Type	Bombing
Perpetrator: Individual ID	urban guerrillas
Perpetrator: Organization ID	FMLN
Perpetrator: Confidence	Suspected or Accused by Authorities: FMLN
Physical Target: Description	vehicle
Physical Target: Effect	Some Damage: vehicle
Human Target: Name	Roberto Garcia Alvarado
Human Target: Description	attorney general: Roberto Garcia Alvarado driver bodyguards
Human Target: Effect	Death: Roberto Garcia Alvarado No Injury: driver Injury: bodyguards

Target Applications

- Converting unstructured texts to databases.
  - E.g. from Wikipedia to DBpedia.
- Providing input to summarization systems.
- Creating indexes for Information Retrieval systems.



DBpedia Blog | Get Involved | Get Help

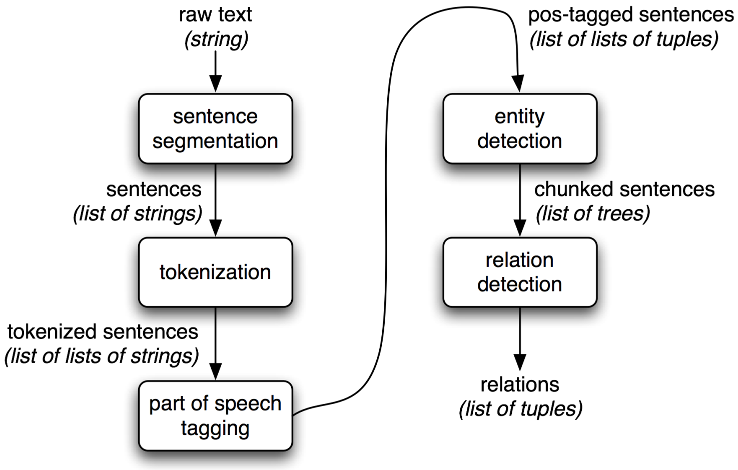
About / News
Applications
Use Cases
Datasets
Online Access
DBpedia Live
Downloads
Interlinking
Development
Support
Publications

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. We hope that this work will make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.

Upcoming Events

2nd DBpedia Community Meeting
Leipzig, September 3, 2014
#DBpediaLeipzig2014
co-located with SEMANTiCS Leipzig 2014

IE Architecture



http://nltk.org/book/ch07.html

Programme

1. Information Extraction
2. Named Entity Recognition
   - Rule-based NER
   - Statistical NER

What are Named Entities?

Named entities are (often multi-word) expressions that refer to proper names of:

- persons,
- organisations,
- locations,
- artifacts,
- dates,
- etc.

Example Text

Text

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.
Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CEO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Example Text

Persons

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.
Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CEO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Example Text

Organisations

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CEO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Example Text

Locations

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CEO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Example Text

Artifacts

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.
Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CEO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Example Text

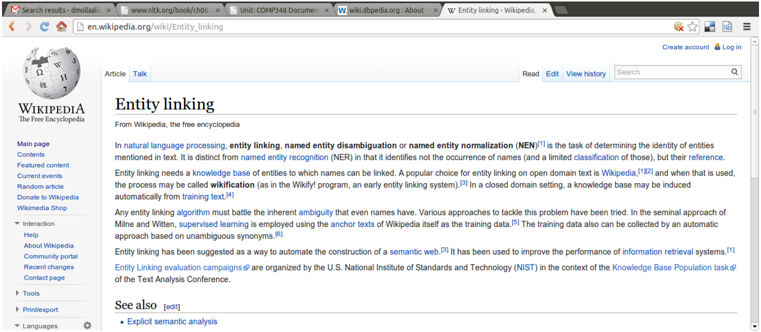
Dates

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CEO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

NER for Entity Linking

- Entity linking is about grounding an expression in a document to a database or to an URI.
- It is a popular means to enrich the contents of webpages.



Entity linking

From Wikipedia, the free encyclopedia

In natural language processing, **entity linking**, **named entity disambiguation** or **named entity normalization** (**NEN**)[1] is the task of determining the identity of entities mentioned in text. It is distinct from named entity recognition (NER) in that it identifies not the occurrence of names (and a limited classification of those), but their reference.

Entity linking needs a knowledge base of entities to which names can be linked. A popular choice for entity linking on open domain text is Wikipedia,[1][2] and when that is used, the process may be called **wikification** (as in the Wikify! program, an early entity linking system).[3] In a closed domain setting, a knowledge base may be induced automatically from training text.[4]

Any entity linking algorithm must battle the inherent ambiguity that even names have. Various approaches to tackle this problem have been tried. In the seminal approach of Milne and Witten, supervised learning is employed using the anchor texts of Wikipedia itself as the training data.[5] The training data also can be collected by an automatic approach based on unambiguous synonyms.[6]

Entity linking has been suggested as a way to automate the construction of a semantic web.[3] It has been used to improve the performance of information retrieval systems.[1]

Entity Linking evaluation campaigns are organized by the U.S. National Institute of Standards and Technology (NIST) in the context of the Knowledge Base Population task of the Text Analysis Conference.

See also [edit]

- Explicit semantic analysis

Issues I

Why not simply using lists of names of people and organisations?

1. It is not possible to list all possible names in the world;
2. new names are formed all the times;
3. names can occur in variations;
4. names of organisations can be complex entities;
5. words are ambiguous.

Examples of variations

- “The Royal Bank of Scotland plc”
- “The Royal Bank of Scotland”
- “The Royal plc”
- “The Royal”
- “RBS”

Issues II

Complex entities with conjunctions

- “China International Trust and Investment Corp”
- “Mason, Daily and Partners”

Overlap between types of names

- “Philip Morris” as an organisation
- “Philip Morris” as a person
- “Washington” as a location
- “Washington” as a person

Other word ambiguities

- “Hope” and “Lost” as proper names (location)
- “Hope” and “Lost” as common nouns

Programme

1. Information Extraction
2. Named Entity Recognition
   - Rule-based NER
   - Statistical NER

Regular Expressions

- Many time and number expressions can be easily handled via regular expressions.
- Need to handle elided elements and referring expressions properly.

Examples

- “Schneider said this year would be his last with the symphony”
- “The contractor submitted a proposal on Tuesday. The day after that, the contract was awarded. That night, they had a party.”

Techniques for Identifying Names

- Entity names:
    - persons, locations, organisations, artifacts, ...
- Source of information:
    - Use context inside the document.
    - Use external knowledge (gazetteers).
- Each of these sources can be exploited along a continuum from cheaper to computationally and manually more expensive usage.

Gazeteers: External Knowledge

- A gazetteer is a geographical dictionary.
- It is an important reference for info about place names.
- It typically contains additional information concerning:
  - The geographical makeup of a country or region.
  - The social statistics and physical features such as mountains, waterways, or roads.

Gazetteers on the Web

- There are many gazetteers or name lists on the web.
- The Global Gazetteer:
  - http://www.fallingrain.com/world
  - A directory of 2,880,532 of the world's cities and towns.
- Tageo.com:
  - http://www.tageo.com/index.htm
  - Provides information about 2,667,417 cities in the world.
- Geographic Names of Australia:
  - http://www.ga.gov.au/place-names/

Information Inside the Document

Internal Evidence

Evidence present in the name itself:

- Corporate designators: “Ltd”, “Inc”, “Pty”.
- Titles: “Mr”, “Dr”, “Rt Hon”.
- Uppercase patterns.
- Special formats (e.g. in numbers and dates).

External Evidence

Evidence given by other parts of the document (usually near the candidate):

- Specific words: “General Motors analyst”.

Programme

1. Information Extraction
2. Named Entity Recognition
   - Rule-based NER
   - Statistical NER

NER as Sequence Labelling

Using Statistical Methods

We can use sequential classifiers like in Part of Speech tagging.

First Attempt

Fletcher/PER Maddox/PER ,/O former/O Dean/O of/O the/O UCSD/ORG Business/ORG School/ORG ,/O announced/O the/O formation/O of/O La/ORG Jolla/ORG Genomatics/ORG together/O with/O his/O two/O sons/O ./O La/ORG Jolla/ORG Genomatics/ORG will release its product Geninfo/ART in/O June/DATE 1999/DATE ./O Geninfo/ART is/O a/O turnkey/O system/O to/O assist/O biotechnology/O researchers/O in/O keeping/O up/O with/O the/O voluminous/O literature/O in/O all/O aspects/O of/O their/O field/O ./O

NER versus PoS

Problems

- How do you differentiate two adjacent NEs of the same kind?
- Rules governing the first word of a NE might be different from those governing following words.

The Approach: IOB Notation

- Every NE category creates two classification tags:

  B Begin of a named entity.

  I In a named entity.

- Any word which is not part of a named entity has a special tag:

  O Outside any entity.

Example

Fletcher/PER-B Maddox/PER-I ,/O former/O Dean/O of/O the/O UCSD/ORG-B Business/ORG-I School/ORG-I ,/O announced/O the/O formation/O of/O La/ORG-B Jolla/ORG-I Genomatics/ORG-I together/O with/O his/O two/O sons/O ./O La/ORG-B Jolla/ORG-I Genomatics/ORG-I will release its product Geninfo/ART-B in/O June/DATE-B 1999/DATE-I ./O Geninfo/ART-B is/O a/O turnkey/O system/O to/O assist/O biotechnology/O researchers/O in/O keeping/O up/O with/O the/O voluminous/O literature/O in/O all/O aspects/O of/O their/O field/O ./O

The General Approach

Training

1. Convert NE annotations into token-based annotations:
   - B, I, O
2. Train the sequential classifier.

Running the NER

1. Convert NE annotations into token-based annotations.
2. Run the sequential classifier.
3. Convert the token-based tags back into multiple-word-based NE labels.

The General Approach

Training

1. Convert NE annotations into token-based annotations:
   - B, I, O
2. Train the sequential classifier.

Running the NER

1. Convert NE annotations into token-based annotations.
2. Run the sequential classifier.
3. Convert the token-based tags back into multiple-word-based NE labels.

Generating the Final Named Entities

What are the Named Entities generated here?

Dr./PER-I Maddox/PER-I will/O be/O the/O firm/O ’s/O CEO/O ./O His/O son/O ,/O Oliver/PER-B ,/O is/O the/O Chief/O Scientist/O and/O holds/O patents/O on/O many/O of/O the/O algorithms/O used/O in/O Geninfo/ART-I ./O

General rules that we used in our AFNER (2008) system

(http://afner.sourceforge.net/)

- A token labelled with B starts a new NE.
- A token labelled with I:
  - Starts a new NE if the previous token belongs to a different entity.
  - Continues the previous entity otherwise.
- A token labelled with O does not belong to any entity.

What Machine Learning Tool to Use?

- Named Entity Recognition is a sequence labelling task.

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix} = \begin{pmatrix} \text{PER-I,} & \text{PER-I,} & \text{O,} & \text{O,} & \text{O,} & \text{O,} & \text{O,} & \text{O} \\ \text{Dr.,} & \text{Maddox,} & \text{will,} & \text{be,} & \text{the,} & \text{firm,} & \text{'s,} & \text{CEO} \end{pmatrix}$$

- In contrast with PoS tagging, however, we may need to use many features.
- In addition, we may need to look at context wider than bigrams.
- HMMs do not model complex features easily.
- Conditional Random Fields are a better choice for this kind of more complex sequence labelling.
- In our solution (AFNER), we used a standard classifier with complex context features.

Features for Classification I

Here are the features used in our own NER called AFNER:

Regular Expressions, Gazetteers

Regular Expressions	Specific patterns for dates, times, etc
FoundInList	The token is a member of a gazetteer

Internal Token Properties

InitCaps	The first letter is a capital letter
AllCaps	The entire word is capitalised
MixedCaps	The word contains upper case and lower case letters
IsSentEnd	The token is an end of sentence character
InitCapPeriod	Starts with capital letter and ends with period
OneCap	The word is a single capitalised letter
ContainDigit	The word contains a digit
NumberString	The word is a number word (‘one’, ‘thousand’, etc.)

Contextual Features

PrepPreceded	The word is preceded by a preposition (in a window of 4 tokens)
PrevClass	The class assigned to the previous token
ProbClass	The probability assigned to a particular class in the previous token

Features for Classification II

Global Features

AlwaysCapped	The token is capitalised every time it appears in the document

Take-home Messages

- Define what is IE, NER.
- Sketch a generic architecture of an IE system.
- Use IOB annotation to mark named entities in a text.
- Sketch the key approaches for rule-based and statistical NER.
- Explain why simply using a list of names will not suffice for NER.
- Compare rule-based and statistical NER approaches.
- Develop regular expressions for simple entities (e.g. numbers, dates).

What's Next

Week 7

- Semantic Web by Rolf Schwitter.
- Friday 27 April, 11pm.