Information Extraction and NER

Natural Language Processing

Master in Business Analytics and Big Data

acastellanos@faculty.ie.edu

Information Extraction

- Information extraction (IE) systems
  - Find and understand **limited relevant** parts of texts
  - Produce a **structured representation** of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
- Goals
  - Organize information so that it is useful to people
  - Put information in a semantically precise form that allows **further inferences to be made** by computer algorithms

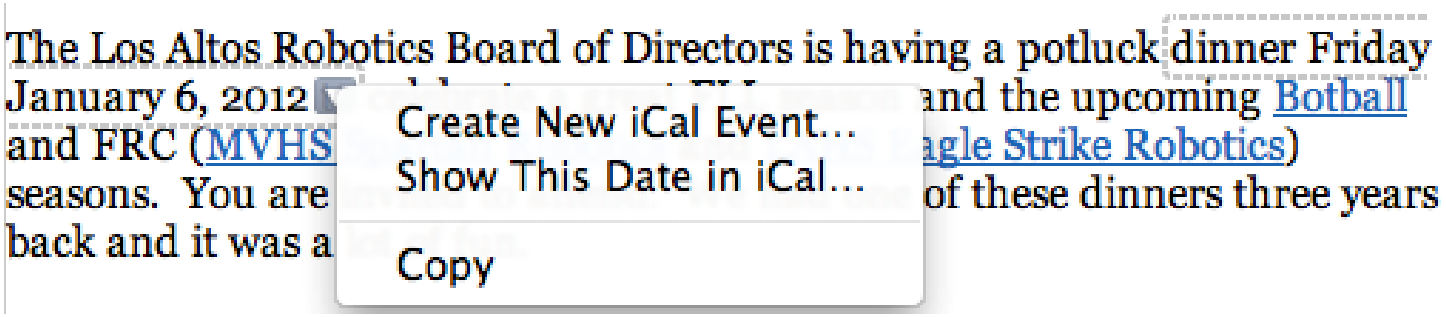
Information Extraction

- Examples
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - `The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.`

      headquarters("BHP Biliton Limited", "Melbourne, Australia")
  - Learn drug-gene product interactions from medical research literature
  - Find symptoms in EHR

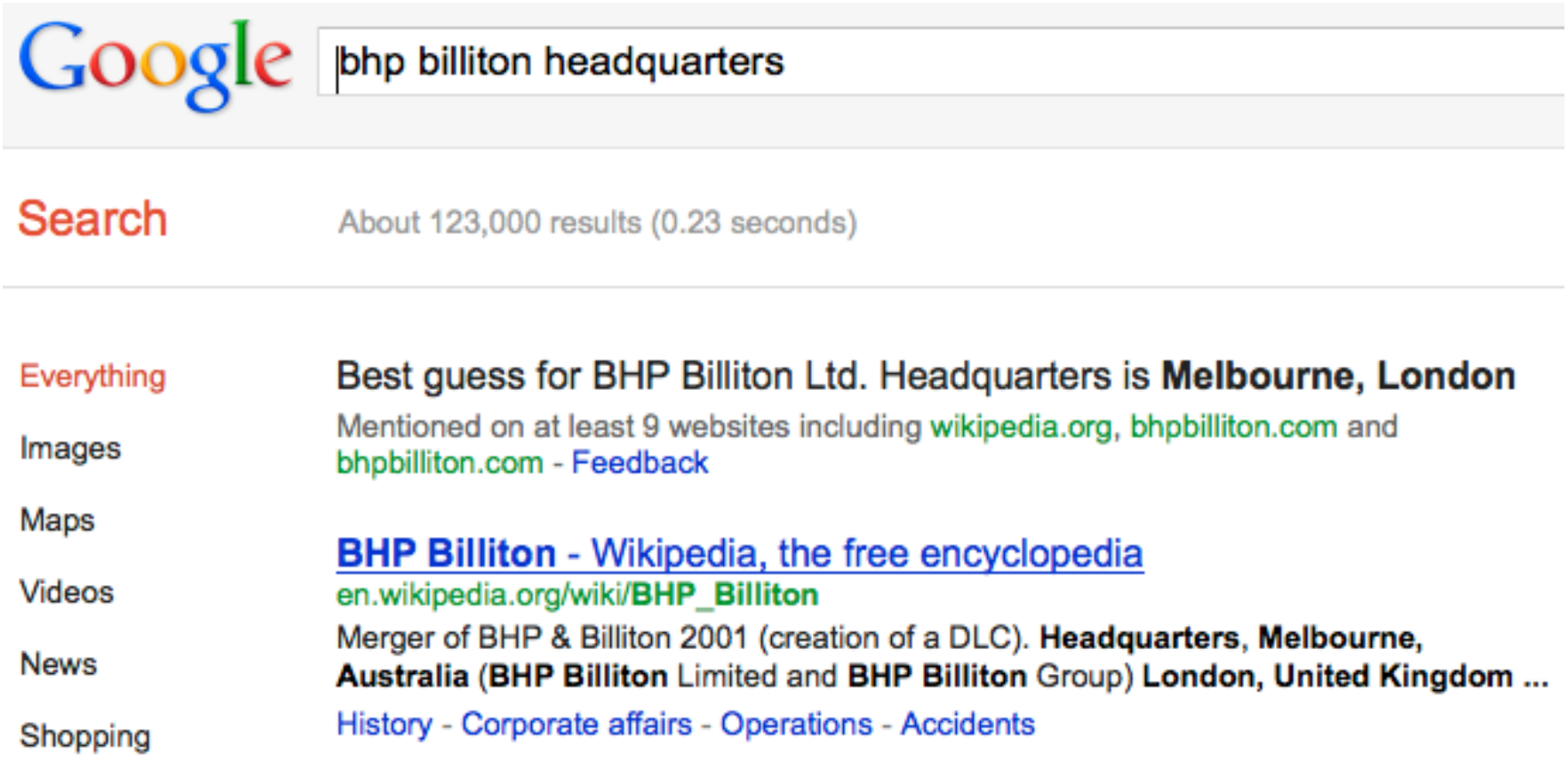
Low-level information extraction

- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing



- Often seems to be based on **regular expressions** and **name lists**

Low-level information extraction



Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:

```
The decision by the independent MP Andrew Wilkie
to withdraw his support for the minority Labor
government sounded dramatic but it should not
further threaten its stability. When, after the
2010 election, Wilkie, Rob Oakeshott, Tony Windsor
and the Greens agreed to support Labor, they gave
just two guarantees: confidence and supply.
```

Named Entity Recognition (NER)

- A very important sub-task: **find** and classify names in text, for example:

The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and **classify** names in text, for example:

The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person
Date
Organization

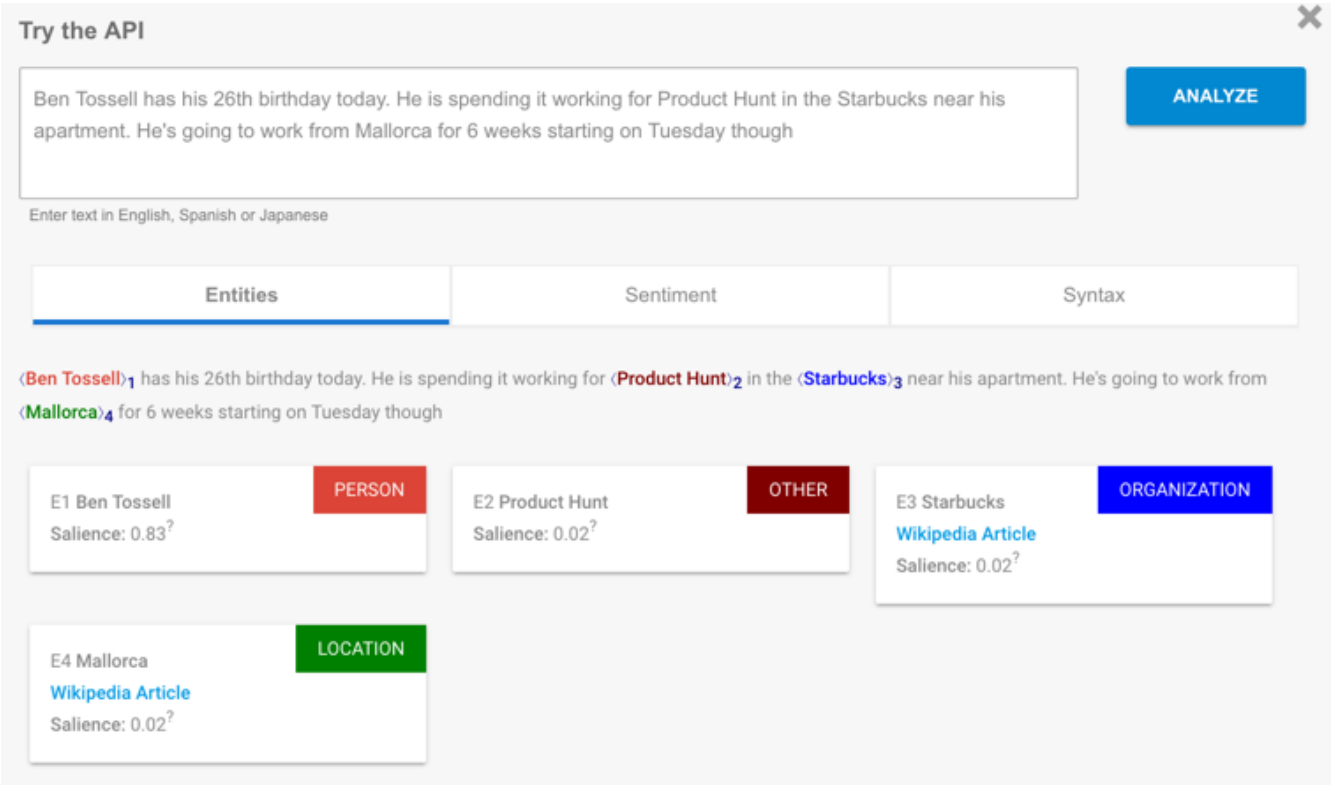
Named Entity Recognition (NER)

- **Sentiment** attributed to companies or products
- A lot of IE relations are associations between named entities
  - **headquarters("BHP Biliton Limited", "Melbourne, Australia")**
- For question answering, answers are often named entities.
  - **who's the president of USA?**

- **Examples:**
  - Many web pages tag various entities, with links to bio or topic pages, etc.
    - Reuters' OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction, …
  - Apple/Google/Microsoft/… smart recognizers for document content

Named Entity Recognition (NER)

- **Cloud Natural Language API**



Named Entity Recognition (NER)

- **Stanford NER**

Please enter your text here:

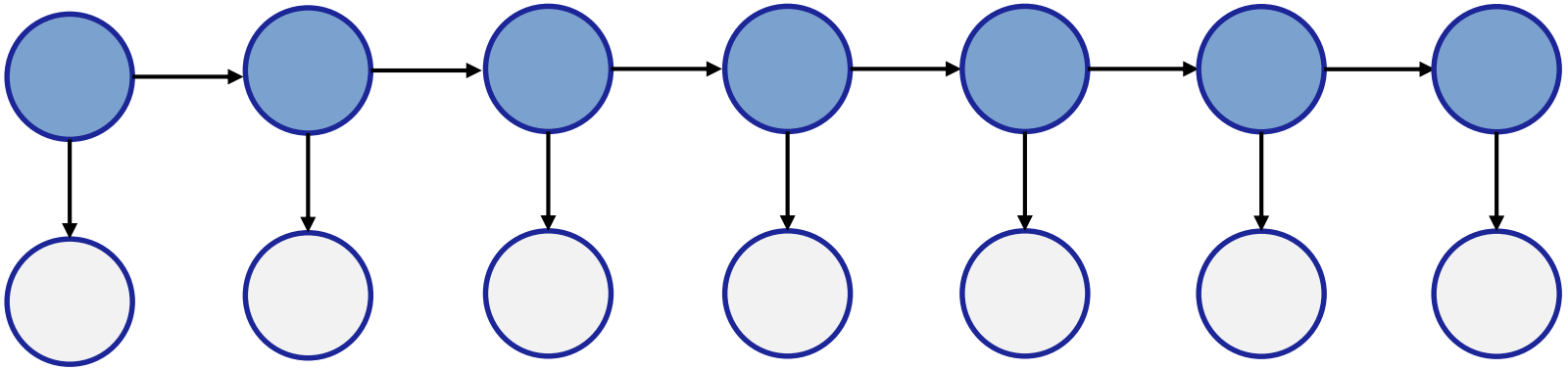
The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Enviar Clear

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

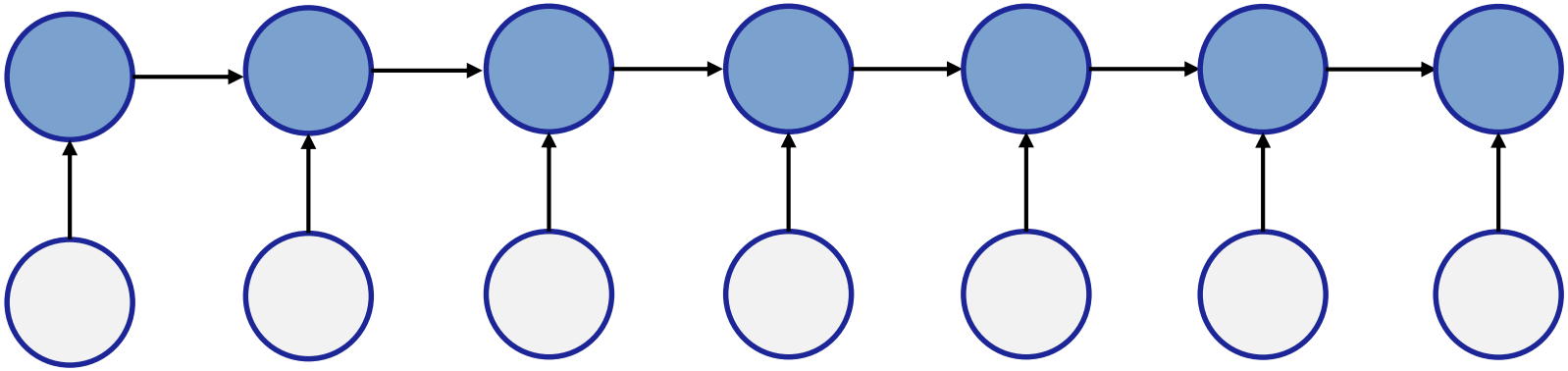
Potential tags:
ORGANIZATION
LOCATION
PERSON

Hidden Markov Models



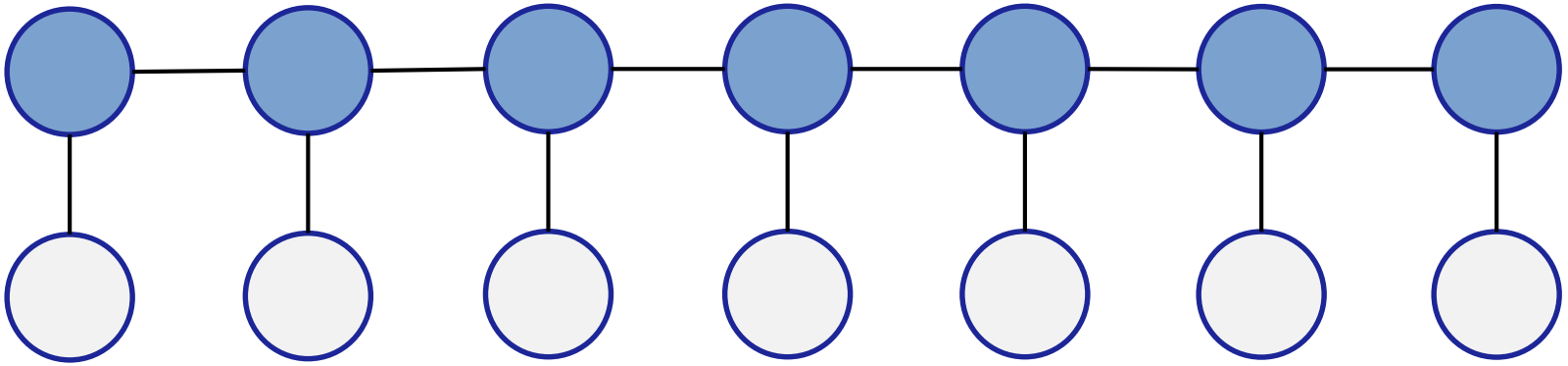
- Generative
  - Find parameters to maximize $P(X, Y)$
- Assumes features are independent
- When labeling X_i future observations are taken into account (forward-backward)

MaxEnt Markov Models



- Discriminative
  - Find parameters to maximize $P(Y|X)$
- No longer assume that features are independent
- Do not take future observations into account (no forward-backward)

Conditional Random Fields



- Discriminative
- Doesn't assume that features are independent
- When labeling Yi future observations are taken into account

The best of both worlds!