

# Information Extraction and NER

Natural Language Processing

Master in Business Analytics and Big Data

[acastellanos@faculty.ie.edu](mailto: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

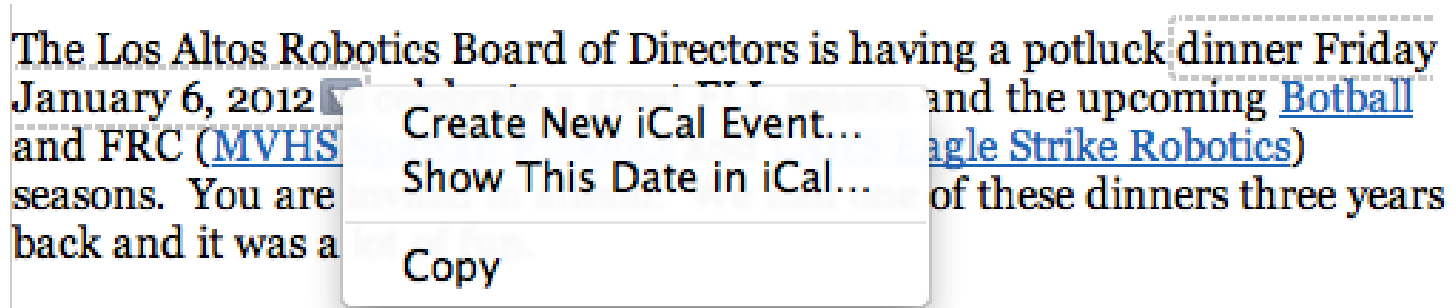
# Information Extraction

## ● 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



bhp billiton headquarters

Search

About 123,000 results (0.23 seconds)

Everything

Best guess for BHP Billiton Ltd. Headquarters is **Melbourne, London**

Images

Mentioned on at least 9 websites including [wikipedia.org](#), [bhpbilliton.com](#) and [bhpbilliton.com](#) - [Feedback](#)

Maps

[BHP Billiton - Wikipedia, the free encyclopedia](#)

Videos

[en.wikipedia.org/wiki/BHP\\_Billiton](#)

News

Merger of BHP & Billiton 2001 (creation of a DLC). **Headquarters, Melbourne, Australia (BHP Billiton Limited and BHP Billiton Group) London, United Kingdom ...**

Shopping

[History - Corporate affairs - Operations - Accidents](#)

# 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

Try the API
×

Ben Tossell has his 26th birthday today. He is spending it working for Product Hunt in the Starbucks near his apartment. He's going to work from Mallorca for 6 weeks starting on Tuesday though

Enter text in English, Spanish or Japanese

Entities

Sentiment

Syntax

Ben Tossell<sub>1</sub> has his 26th birthday today. He is spending it working for Product Hunt<sub>2</sub> in the Starbucks<sub>3</sub> near his apartment. He's going to work from Mallorca<sub>4</sub> for 6 weeks starting on Tuesday though

E1 Ben Tossell

Saliency: 0.83<sup>?</sup>

PERSON

E2 Product Hunt

Saliency: 0.02<sup>?</sup>

OTHER

E3 Starbucks

[Wikipedia Article](#)

Saliency: 0.02<sup>?</sup>

ORGANIZATION

E4 Mallorca

[Wikipedia Article](#)

Saliency: 0.02<sup>?</sup>

LOCATION

# 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.

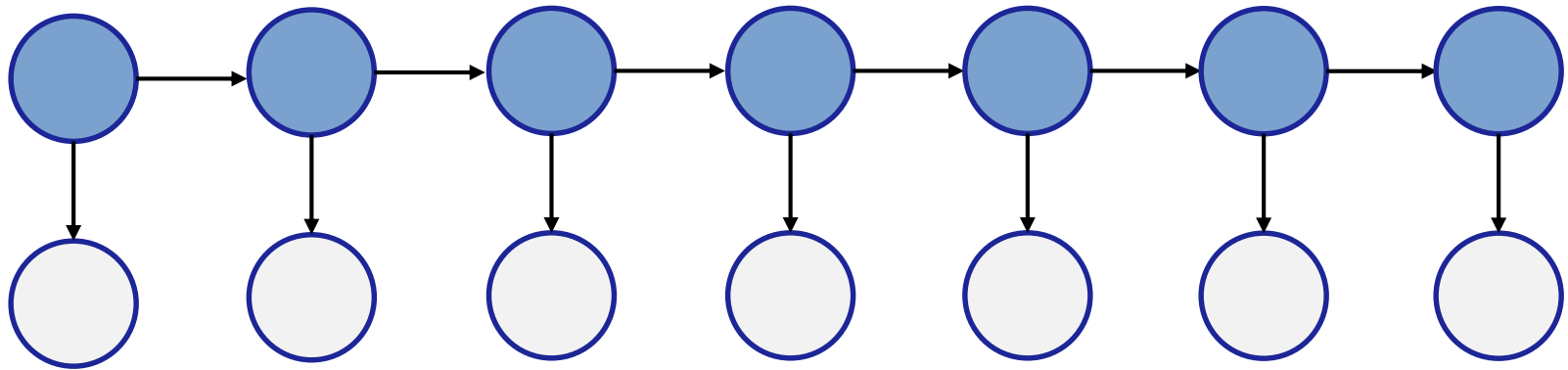
Potential tags:

**ORGANIZATION**

**LOCATION**

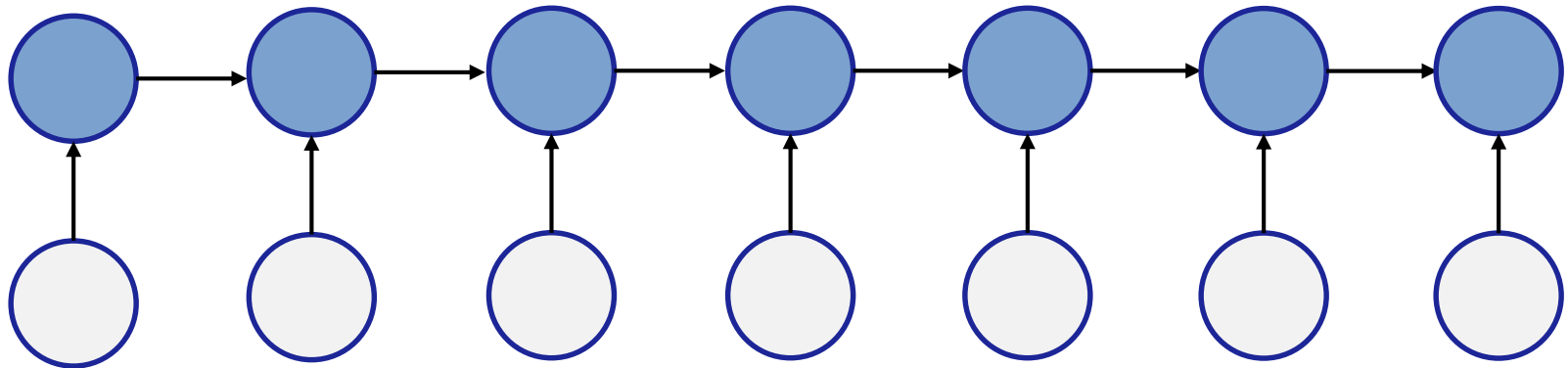
**PERSON**

# Hidden Markov Models



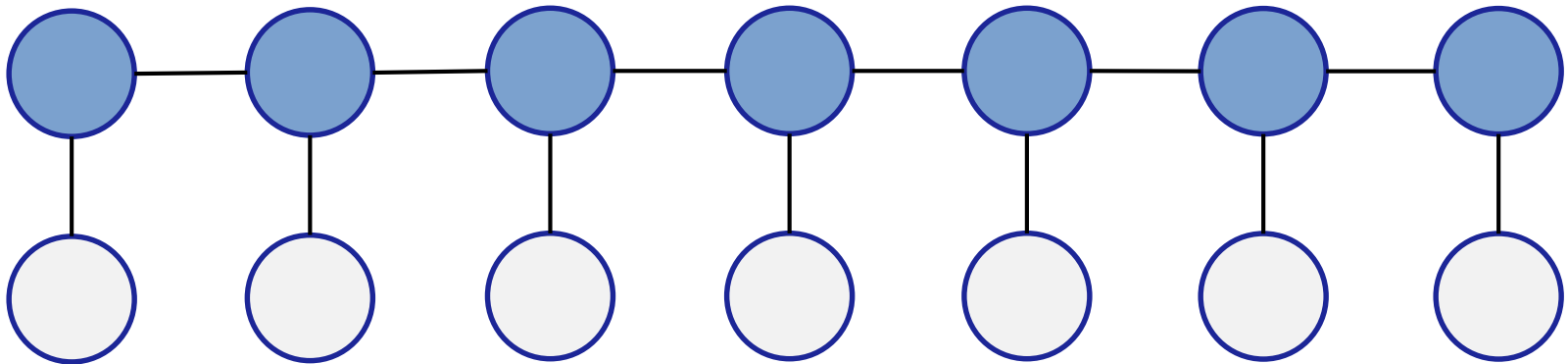
- Generative
  - Find parameters to maximize  $P(X, Y)$
- Assumes features are independent
- When labeling  $X_i$  future observations are taken into account (forward-backward)

# 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  $Y_i$  future observations are taken into account

**The best of both worlds!**