

Information Extraction and Named Entity Recognition

Introducing the tasks:
Getting simple structured
information out of text



Information Extraction

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



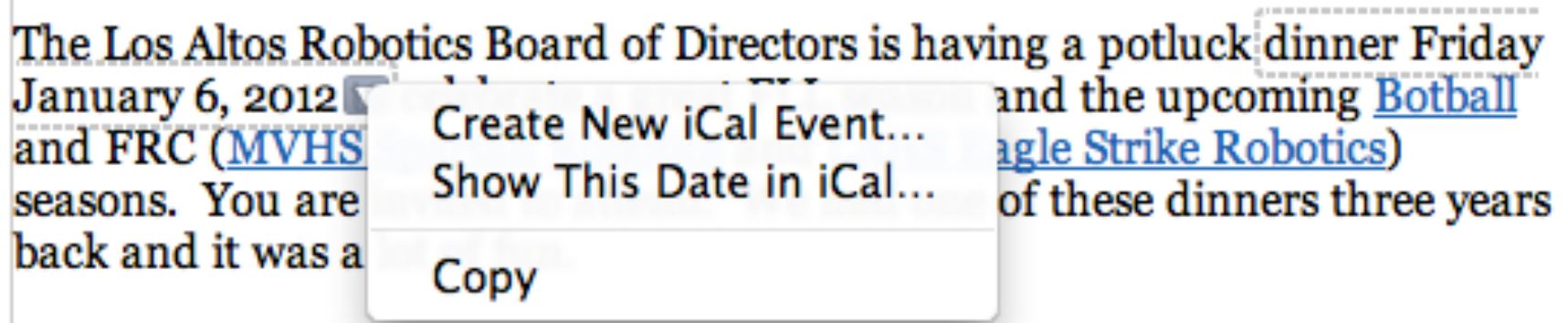
Information Extraction (IE)

- IE systems extract clear, factual information
 - Roughly: *Who did what to whom when?*
- E.g.,
 - 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



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](http://en.wikipedia.org/wiki/BHP_Billiton)

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 |
| Location |
| Organi- zation |



Named Entity Recognition (NER)

- The uses:
 - Named entities can be indexed, linked off, etc.
 - Sentiment can be attributed to companies or products
 - A lot of IE relations are associations between named entities
 - For question answering, answers are often named entities.
- Concretely:
 - 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



Information Extraction and Named Entity Recognition

Introducing the tasks:
Getting simple structured
information out of text