

# Concept identification and Processing of multi-document discourse

David Martinez



Language Technology Group

The University of Melbourne

<http://www.cs.mu.oz.au/research/lt/>

# Concept identification

- My research background on Word Sense Disambiguation (WSD).
- Attempts to integrate this technology for search engines limited success so far.
- However, critical to search for relevant information in some highly specialised domains. E.g. Biomedicine.

# Concept identification

- Project with the University of Sheffield on WSD for the Biomedical domain.
- Motivation: NLM (National Library of Medicine) Indexing Initiative to link texts to ontology.
  - Indexes 3,700 citations a day, 5 nights a week.
  - Users goal is to find very specific information that is difficult to reach with keyword matching.
- High polisemy in texts:
  - Acronyms
  - Regular polisemy (e.g. gene/protein)
  - Mix of general/specific uses (e.g. *was* as a gene or a verb).

# Concept identification

- LT tools can help, from PoS taggers to supervised WSD.
- Resources and tools being developed in this domain.
  - Metathesaurus
  - Gene ontologies
  - Medical WordNet
- Applications: terminology recognition, gene classification, knowledge discovery, indexing.

## Processing of multi-document discourse

- Discovery project at the University of Melbourne: Information Delivery
- Main challenge for search technologies is to reach information from multi-document discourse (e.g. web forums, mailing lists)
- Rich and dynamic source of information.
- Latest information provided by this medium.
- Difficult for new users to catch up.

# Processing of multi-document discourse

- Knowledge is not accessible with shallow techniques.
- Multi-layered analysis of text required to acquire structured data:
  - Discourse analysis
  - Thread classification
  - Information Extraction: concepts and relations
  - Rich linguistic information

# Processing of multi-document discourse

- Our approach:
  - Whiteboard: Integration of different processors in common architecture.
  - Precision grammars: augment coverage with deep lexical acquisition.
  - Specific domain: Linux newsgroups
  - Application: Web service to deliver acquired information in structured way.

# Concept identification and Processing of multi-document discourse

David Martinez



Language Technology Group

The University of Melbourne

<http://www.cs.mu.oz.au/research/lt/>