Semantik VT07
Semantiska och pragmatiska tillämpningar
Stina Ericsson
(med tack till Torbjörn Lager för många oh-bilder)

Semantics in NLP applications
- Machine translation
- Natural language generation
- Automatic summarisation
- Dialogue systems
- Text categorisation
- Spam filtering
- Context-based navigation
- Information extraction
- Natural language search
- Question-Answering
- Grammar checking
- Computer-assisted language learning
- Spell checking and correction

Semantics in machine translation

Machine translation systems
- SYSTRAN is behind:
  - http://babelfish.altavista.com/
  - Google’s and Yahoo’s online translation systems

Semantics and translation
- Direct translation, Transfer approach, Interlingua approach
- Analyse source text
- interlingua transfer
- direct translation
- Target text

Challenges for machine translation
- Polysemy, synonymy, homonymy
  - Machine translation needs WSD
- Syntactical ambiguities
- Referential ambiguities
- Metaphors and symbols
- New vocabulary development
Compositionality in translation

Basic idea in machine translation:
The translation of a complex expression is determined by the translation of its parts, and the way in which those parts are combined.

Semantics in IR, IE and QA

Semantics in Information Retrieval

Query expansion
If a relevant document does not contain the terms that are in the query, then that document will not be retrieved. The aim of query expansion is to reduce this query/document mismatch by expanding the query using words or phrases with similar meaning.

Semantics in Information Retrieval

IR needs WSD
When searching the web for information about companies buying shares in other companies, we don’t want to retrieve information about interest rates.

Two kinds of Information Extraction

Closed-domain IE
- Info about joint ventures from business news
- Understanding weather reports
- Summarising stock market reports
- Summarising terrorist reports

Open-domain IE

Approaches to IE

‘Deep’ methods vs. ‘shallow’ methods
- Tacitus (deep)
- Proteus (in between)
- LaSIE-II (fairly shallow)
- Fastus (shallow)

Manual knowledge engineering vs. machine learning
The LaSIE-II system (Sheffield): modules

- Tokeniser
- Gazetteer lookup
- Sentence splitter
- Brill tagger
- Morphological analyser
- Parser
- with named-entity grammar
- with general phrasal grammar
- Name matcher
- Discourse interpreter
- Template writer

Levels of processing in FASTUS

- Tokens
- Complex words (multi-word phrases, number, proper names)
- Basic phrases
- Complex phrases
- Semantic patterns (identify entities and events)
- Merging (merge references to the same entity or event from different parts of the text)

Why open-domain Q&A needs WSD

ID 29: Show me info about the proposed development plans for the south bank of Thames.
ID 16: Information about computerised patient records in Australia.
ID 44: Give me info about the performance of property brokers in New Zealand.
ID 33: tell me more about this platinum mine in South Africa.
ID 27: Looking for recent articles on insurance companies interested in game theory or other new economic models.
ID 60: I want information in English about the capacities of chemical plants.
ID 28: Show me information about the gas agreement that Pakistan entered recently.

In ID29, the CCS code for 'BusinessSect0r:Banking' is wrongly introduced
The word "record" in ID16 triggers the CCS code for 'BusinessSect0r:Disco/Tapes'
The word "performance" in ID44 introduces the code for 'NewsEventType:Performance or exhibition', which is wrong
The word "mine" in ID33 introduces the code for 'BusinessSect0r:Ammunition'
The word "game" in ID27 triggers the code for 'BusinessSect0r:Toys and Games'
The word "plants" in ID60 triggers the code for the botanical concept of plant
The word "gas" in ID28 triggers the code for 'BusinessSect0r:Refined Oil', probably since "gas" is short for "gasoline" in American English.

MUC (Message Understanding Conference)

IE: "the next step up from search engines in fulfilling information processing needs"
- Named entities + attributes, relations, events

A web application

- http://www.ask.com/

Semantics in dialogue systems
Semantics in dialogue systems

- Natural language interfaces to databases
- Conversational agents
- Real dialogue systems

Conversational agents

- Parry (Colby, 1972) http://en.wikipedia.org/wiki/PARRY
- The Turing test
- The Loebner prize http://www.loebner.net/Prizef/loebner-prize.html

Web agents

- Hanna http://www.fk.se/privatpers/
- Sandra http://www.tullverket.se/
- Anna http://www.ikea.se

VoiceXML

- An XML-based language for the development of telephone-based dialogue systems
- A mark-up language
- Standard developed by W3C
- Semantics in VoiceXML
  - Attribute-value matrices

VoiceXML example:

```xml
<form id="traveling">
  <grammar src="travel.gxml" type="application/xgml"/>
  <initial name="get_info">
    <prompt> How can I help you? </prompt>
    <catch event="nomatch">
      <prompt> Let's try getting each field separately. </prompt>
      <prompt> <assign name="get_info" expr="true"/>
    </prompt>
  </initial>
  <field name="from_city">
    <grammar src="travel.gram1#fromWhoa"/>
    <prompt> What city are you travelling from? </prompt>
  </field>
  <field name="to_city">
    <grammar src="travel.gram1#toWherea"/>
    <prompt> What city are you travelling to? </prompt>
  </field>
  <block>
    <submit next="http://www.example.com/servlet/flipway"/>
  </block>
</form>
```
Some Swedish dialogue systems

DJGoDIS, Göteborg
AdApt at KTH, Stockholm
August, KTH

The Semantic Web

http://www.w3.org/2001/sw/
Provides a common framework that allows data to be shared and reused across applications, etc
W3C + researchers and industry
OWL (Web Ontology Language)
- for publication and sharing of ontologies
RDF (Resource Description Framework)
- to represent information and exchange knowledge

Text summarisation

A web application

SweSum
http://swesum.nada.kth.se