Language engineering systems – seminar 1

- Information access
- Information retrieval
- Text summarization
- Question answering
- Information extraction
- Presentations
  - Open-domain question answering techniques (Andreas)
  - Information extraction, Automatic (Åsa)
- Evaluation (slides from lecture 1)

Information retrieval and language technology

- Information retrieval
  - Retrieving documents in response to queries
  - Documents represented as word (or term) vectors
  - Words represented as document vectors
- LT in IR
  - Morphological analysis often does not improve on stemming
  - Compounding and inflection forms might be important in keyword based search engines

Information retrieval and language technology

- IR in LT
  - Latent semantic indexing (LSI) and the distributional theory of meaning.
  - Vector-space models have been applied to a number of LT tasks such as word-sense disambiguation and question classification.
- Evaluation methodology
  - Experts creating gold standards.
  - Measures such as recall and precision
Information retrieval and language technology

- LT extensions
  - Automatic summarization
    - Judging relevance
      - "What is the main message of this document?"
    - Question answering
      - Finding answers in documents to specific questions, such as "What is the capital of Honduras? When was Albert Einstein born?" (and more complex ones)
    - Information extraction
      - "What information is given in these documents on events of a given, specified type?"

Text summarization

- Extraction approach
  - Identify the sentences of a document that carry the most important information in the document and join them to a paragraph.
  - Features to look at:
    - Co-reference problems
    - Long vs short summaries
    - Generically applicable method

- Abstraction approach
  - Create a connected representation of the contents of the text and generate a text from that.
  - Requires parsing and semantic interpretation
  - Domain-specific modelling required

Text summarization evaluation

- Intrinsic evaluation ("Is this a good summary?")
  - Criteria: coherence vs informativeness
  - Based on gold standards / reference summaries
  - Automatic scoring
  - Human evaluation comparing text and summary

- Extrinsic evaluation ("Is this a useful summary for some task?")
  - Criteria and measures:
    - E.g. human understanding as measured by reading comprehension tests
    - Human task efficiency as measured by relevance judgements