The future of corpus linguistics

Martin Volk
Universität Zürich

Topics

- Corpus types and size
- Corpus annotation
- Corpus access
- NLP Applications

Future corpus annotation

- Ever larger automatically annotated corpora
- Manually annotated corpora
  - grow in depth (additional annotation)
  - grow in breadth (additional data)
  - grow across languages

Future corpus annotation

- Text coherence annotation
  - coreferences (across sentences)
- Semantic annotation
  - name classes
  - local, temporal, modal, causal units
  - word senses relative to a thesaurus
  - roles within a sentence (FrameNet)
    (cp. to http://www.icsi.berkeley.edu/~framenet/)

Future corpus annotation

- Parallel treebanks
  - semi-manual annotation of translated texts
  - alignment on word and constituent level
  - the parallel text may serve as disambiguator

Future corpus annotation

- Better parsers for corpus annotation
  - integration of shallow parsing and deep parsing
- Long term:
  - automatic transcription of spoken language data
  - automatic search through audio- and video data

Future corpus access

- through the web
- the web as corpus
- to multilingual corpora through the alignment of parallel texts

Future applications

- From information retrieval to answer retrieval
- From document retrieval to fact retrieval
- Information Gathering and Automatic Summarization
- Document Classification
- Dialogue Systems ("lingubots")
  - automatic email answering
  - integration of spoken and written language with graphics and images
- E-Learning

Observations

- The Translation-Memory lesson: A sentence S is best translated by retrieving a human translation T. (A human translation is nothing but an annotated corpus.)
  - The challenge: Match all variants of S.
- The FAQ lesson: A question Q is best answered by finding the Q+A pair in the database.
  - The challenge: Find all variants of Q.

Conclusions from the Observations

- A lot can be gained from systematically harvesting human annotations.
- My prediction: Lexical entries in NLP systems will be replaced by phrases.
- NLP moves **from** detailed analysis of the input and aggregation of the output **to** matching of similar cases (of the input) and adaptation of the output.