Methods and Tools for Automatic Grammar Extraction (GramEx)

Anna Sågvall Hein
Joakim Nivre
Beáta Bandmann Megyesi

Uppsala University
Department of Linguistics and Philology
Introduction

- Project funded by Swedish Research Council 2006–2008
- Three ideas:
  - Grammar extraction from large corpora
  - Evaluated in machine translation
  - By-product: Swedish treebank
Grammar Extraction

- Narrow sense:
  - Use corpus data to induce a formal grammar that can be used for parsing (and/or generation)

- Wide sense:
  - Use corpus data to induce models that can be used for syntactic analysis and generation:
    - Train a statistical parser
    - Train a statistical disambiguator for a grammar-based parser
    - Supervised and unsupervised learning
    - Deductive and inductive learning
Evaluation in MT

- **Narrow sense:**
  - Use extracted grammar for parsing in MT; compare to original hand-crafted grammar

- **Wide sense:**
  - Use induced syntactic model to improve MT:
    - Parsing for syntax-based (statistical) MT
    - Parsing for improved word (and phrase) alignment
    - Syntax-based translation models (tree transduction)
    - Generation for syntax-based (statistical) MT
Swedish Treebank

- **Narrow sense:**
  - Manually validated syntactic annotation of Swedish corpus

- **Wide sense:**
  - Syntactically annotated corpus data:
    - Semi-automatically annotated Swedish corpus
    - Swedish-Turkish parallel trebank (with or without alignment)
Planning (Fall 2006)

- Establish basic data sets (and tools)
- Establish evaluation framework
- Establish treebank core
Data Sets and Tools

Data selection:
- Talbanken
- SUC
- Parole
- Scarrie
- Swedish-Turkish parallel corpus
- ...

Segmentation:
- Tokenization
- Sentence segmentation

Annotation:
- Part-of-speech tagging
- Morphological analysis (?)
- Lemmatization (?)
- Named entity annotation (?)
- Parsing – multiple approaches
Evaluation and Treebank

- Evaluation frameworks:
  - Parser evaluation using treebank core
  - MT evaluation

- Treebank core:
  - Gold standard annotation for all parsing standards
  - Stratified sample of free gold standard resources?