Interaction, Domain Adaptation and Retraining

Andre Blessing, Kerstin Eckart, Jonas Kuhn, Jens Stegmann, Heike Zinsmeister

Institute for Natural Language Processing (IMS), University of Stuttgart
{andre.blessing|jonas.kuhn|jens.stegmann}@ims.uni-stuttgart.de
CLARIN-D Centre Stuttgart

- Institute for Natural Language Processing (IMS), University of Stuttgart, Germany.
- Foundations of Computational Linguistics Group led by Prof. Dr. Jonas Kuhn
- Aiming for type B status
- Resources and expertise
  - annotated text and speech corpora
  - lexical resources
  - standardization and metadata
  - tools for processing/annotating data
  - web services
  - domain adaptation
- **Problem:**
  - Researchers from the Digital Humanities may want to apply statistically-based NLP tools to very specific textual domains
  - But: the pertinent tools have been trained (and tested) on a particular kind of data (→ newspaper texts from the 80s/90s)
  - Hence, much too often the tools will tend to deliver rather disappointing results
- The expert user could, in principle, give valuable feedback.
- We would like to 'exploit' that users knowledge in order to adapt our tools to new domains.
- That would be a win-win situation:
  → the user would end up with better results
  → the tool creators would end up with more & better models
- But: the means of using web service processing chains/results are rather static at the moment.
- Domain adaptation still has to be brought about.
• Interaction
  - We need nice GUIs and well-tailored modes of visualization for the results of processing
  - They should be interactive, allowing the user to explore, and enable him to
    • mark mistakes,
    • do the corrections, and
    • initiate retraining of the tool.

• Domain Adaptation via Retraining
Interaction

Domain Adaptation via Retraining
  - Retraining
    - Full model retraining
    - Incremental retraining
  - Machine learning techniques
    - Semi-supervised learning
    - Supervised learning
      → Active learning
The IMS-ADAPT Web Application

Annotation web services:
- corpora
- sentence splitting
- tokenizing
- tagging
- morphological analysis
- dependency parsing
- named entity recognition

Classification web service:
- candidate selection
- feature extraction
- training
- rating
- classifier

Visualization:
- annotate instances
- rate extracted instances
- input of free text
An Illustrative Example

- [http://clarin01.ims.uni-stuttgart.de/demo/videos/](http://clarin01.ims.uni-stuttgart.de/demo/videos/)
There are problems galore

- Integrating everything in the repositories
- Making use of means of the infrastructure
- Modularization of tools and models in terms of WSes
- Reloading models can be extremely costly
- Doing everything in parallel for many users

→ Nevertheless: Anybody with us? :)

Thank you for your attention!