Projects: Domain Adaptation for Parsing

Caroline Sporleder & Ines Rehbein

WS 09/10

Domain Adaptation for Parsing

- **Task**: adapt a statistical parser to a new domain
- **Idea**: select/modify training data to make it more similar to the target domain

- 3 Subprojects:
  1. **Clustering of training data**:
     subdivide Penn-II treebank into different domains
  2. **Similarity in space**:
     find domain-specific properties for source and target domain
  3. **Adaptation**:
     train parser on domain-specific training sets

Domain Adaptation for Parsing

- **Task**: adapt a statistical parser to a new domain
- **Idea**: select/modify training data to make it more similar to the target domain

- 3 Subprojects:
  1. **Clustering of training data**:
     subdivide Penn-II treebank into different domains
  2. **Similarity in space**:
     find domain-specific properties for source and target domain
  3. **Adaptation**:
     train parser on domain-specific training sets

Clustering of training data

- Find typical features for instances in the training data
  - POS tags
  - syntax
  - lexical context
  - (POS) n-grams
  - most frequent function words
  - ...
- Cluster sentences according to these features
  $\Rightarrow$ Result: domain-specific subsets for training

Similarity in space

- Once we have domain-specific training sets...
  ... find training set most similar to new domain

  1. Find typical representations/feature sets for text in the target domain
     - ⋆ Attention: raw text, no (gold) syntax available!
  2. Select the subset in the training data which is most similar to the new domain
     - ⋆ use PCA / correlation measures to identify relevant features

Similarity in space

- Once we have domain-specific training sets...
  ... find training set most similar to new domain

  1. Find typical representations/feature sets for text in the target domain
     - ⋆ Attention: raw text, no (gold) syntax available!
  2. Select the subset in the training data which is most similar to the new domain
     - ⋆ use PCA / correlation measures to identify relevant features

Similarity in space

- Once we have domain-specific training sets...
  ... find training set most similar to new domain

  1. Find typical representations/feature sets for text in the target domain
     - ★ Attention: raw text, no (gold) syntax available!
  2. Select the subset in the training data which is most similar to the new domain
     - ★ use PCA / correlation measures to identify relevant features

Adaptation

- Adapt the parser to the new domain by re-training on the domain-specific data set
- Apply further treebank transformations to make training data more similar to target domain, e.g.
  - target domain has less ADJ than source domain
    $\Rightarrow$ delete ADJ from source data
- Do instance weighting on training instances with features which show high correlation to target domain, e.g.
  - give high weights to questions

End product

- 3 modules $\longrightarrow$ 1 project
  - combine into one tool?
  - pipeline architecture?
- approximately 6 students (2 per module)
- need to cooperate / coordinate
  - input/output for each module?
  - operating system / programming language?