Problem: Given training data from several source domains. What data should we use to train a parser for a new domain?

Aim: Find subset of most “similar” articles

Research Questions:

Q1: Which similarity measure is good for parsing?
Q2: How does it compare to manual selection?
Q3: Is it also useful for other languages and/or tasks?

Motivation

- Domain adaptation (DA): adapt system trained on A to work better on B
- Assumption: have data (labeled/unlabeled) available for new domain B

Approach

- We use articles as units (corpus not monolithic)
- Task: Given is a collection of annotated articles, and a new article that we want to parse. Select the most similar articles to train the best parser for that new article.

Measuring Domain Similarity

Similarity Functions

- Jensen-Shannon, Skew divergence
- Euclidean, Variational, Cosine

Feature Representations

We use the simplest representation possible (= just words):

Tools

- MST parser (McDonald et al., 2005)
- MALLET Topic Model toolkit (standard settings); 100 topics; no stopwords removed

Data

- Penn TB Wall Street Journal (WSJ)
- Genia (G) and Brown (B)
- Dutch: larger data set (cf. paper)

Experiment I: Within WSJ

- Human-annotated meta-data available
- Test set: 22 articles from WSJ Section 23 and 24
- Run data selection methods for each article for increasing amounts of data
- Baseline: Random selection
- Compare to selection based on meta-data

Comparison Human-annotated data

- Automatic data selection is better than random and taking all

Experiment II: Domain Adaptation

- Automatic data selection is better than random and taking all

Conclusions & Future Work

- Q1: A simple unsupervised technique (using topic model, closely followed by plain words) is effective for training data selection for parsing (no smoothing/weighting/optimization!)
- Q2: Human-annotated labels did not work better
- Q3: Similarity measure also effective for Dutch and PoS tagging (results in paper)
- Future Work: Evaluate effect with automatically labeled data (self-/up-training), analyze differences between data, data shift detection (Dredze et al., 2010)