Effective Measures of Domain Similarity for Parsing

Barbara Plank, Gertjan van Noord b.plank@rug.nl, vannoord@let.rug.nl University of Groningen, The Netherlands

university of groningen

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?

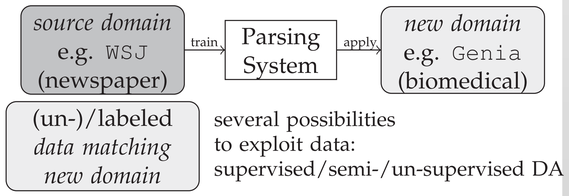
Q3: Is it also useful for other languages and/or tasks?

Result:

- A simple measure based on words is surprisingly effective
- Works better than baseline and human-annotated data
- Less (more specific) data is often better than taking all

Motivation

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



- What if we don't know the target domain? Or there are several source domains?
- **New task/challenge:** Automatic Domain Adaptation (McClosky et al. 2010)

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

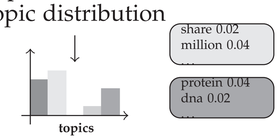
Similarity Functions

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

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

Words:
relative frequencies

Topic model:
topic distribution



Experiments

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)

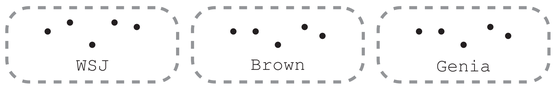
	EN: WSJ	WSJ+G+B	Dutch
articles	2,034	3,776	51,454
sentences	43,117	77,422	1,663,032
words	1,051,997	1,784,543	20,953,850

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

Experiment II: Domain Adaptation

Multiple domains: add Brown and Genia articles



Disregard corpora boundaries. Compare to:

- per-corpus model (knows domain/boundary)
- model trained on all data (union)

Results

Experiment I: Within WSJ

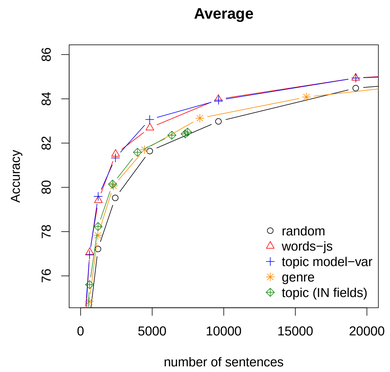
Comparison Similarity Measures

Parsing Accuracy for increasing amounts of training data	1%	3%	12%	25%	97%
random baseline	70.61	77.21	81.64	82.98	85.51
words (Jensen-Shannon)	74.07⋆	79.41⋆	82.69⋆	**83.98**⋆	85.68
topic model (Variational)	**74.29**⋆	**79.59**⋆	**83.06**⋆	83.93⋆	85.43

⋆significantly better than random

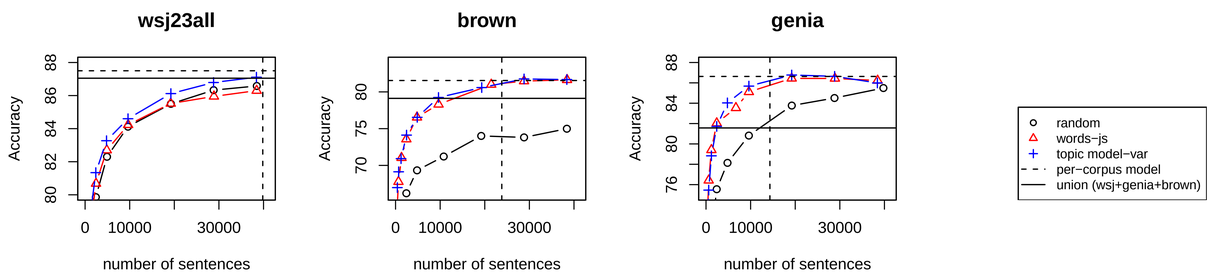
- Automatic measures better than random
- Similarity measures gave similar results → Overlap?
- **Average Overlap** (proportion of identically selected articles): low! (less than 50%)

Comparison Human-annotated data



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)