Introduction and Motivation

- Task: Parsing Natural Language
- Problem: Lack of portability to new domains → drop in performance (Gildea, 2001)
- Examine performance of different dependency parsers for Dutch across Wikipedia domains
  - A grammar-driven (Alpino) versus two data-driven (MST and Malt) parsing systems

Research Question

- How does parser performance for Dutch correlate with simple statistical properties of the text (e.g., average sentence length, unknown word ratio, e.t.c.)?
- First step towards: How is sensitive a given system to the domain, i.e. which system (hand-crafted versus data-driven) is more affected by domain shifts?

Related Work

  - How parser performance correlates to simple statistical measures of the text
- They looked at English (WSJ-Brown), examined several parsers (constituent, dependency, deep-grammar based)
- Build a regression model to predict parser accuracy (WU-Brown)

Parsing Systems

Alpino
- Hand-crafted grammar (HPG-like)
- Separate statistical disambiguation
- Tailored to Dutch

MST (McDonald et al. 2005)
- Data-driven
- Graph-based dependency parser

Malt (Nivre et al. 2007)
- Data-driven
- Transition-based dependency parser

Datasets and Treebank conversion

- Train data - Source: Alpino Treebank (cdb)
  - 7,136 sentences from the Eindhoven corpus (cdb)
  - Average sentence length (ASL): 19.7
  - Collection of test fragments from 6 Dutch newspapers
- Test data - Target: Wikipedia articles
  - 95 Dutch Wikipedia articles which were annotated in the course of the LASSY project
  - Mostly about Belgium issues, i.e. locations, politics, sports, arts, etc.
- We have grouped them into 10 subdomains:

<table>
<thead>
<tr>
<th>Subdomain</th>
<th>Articles</th>
<th>Pages</th>
<th>Tokens</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpino</td>
<td>50</td>
<td>50</td>
<td>1,000,000</td>
<td>500,000</td>
</tr>
<tr>
<td>ASL</td>
<td>50</td>
<td>50</td>
<td>1,000,000</td>
<td>500,000</td>
</tr>
<tr>
<td>LAS</td>
<td>50</td>
<td>50</td>
<td>1,000,000</td>
<td>500,000</td>
</tr>
<tr>
<td>MST</td>
<td>50</td>
<td>50</td>
<td>1,000,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Malt</td>
<td>50</td>
<td>50</td>
<td>1,000,000</td>
<td>500,000</td>
</tr>
</tbody>
</table>

Table: Overview Wikipedia subdomains with associated articles

- Additional (not in paper): DPC corpus
- Dutch Parallel Corpus
- 186 articles from several domains (a.o., medical, oceanography, etc.)
- Conversion Alpino XML to CoNLL format
- Adapted E. Manis's software (CoNLL 2000); retag data with more fine-grained Alpino tags: positive effect
- Retagged data available at: http://www.let.rug.nl/bplank/Alpino2xml
- Simplifications (CoNLL): e.g., tokens have just a single head. MULs are a single token

Statistical Measures and Evaluation

Start from statistical measures used by Zhang & Wang (2009), add perplexity:

1. Average sentence length (ASL)
2. Unknown word rate (UWR):
3. Unknown Part-of-Speech trigram ratio (UPTR):
4. Perplexity:
5. Trigram Language Model perplexity, estimated from cdb corpus

Evaluation

- All parsers are evaluated by Labeled Attachment Score (LAS):
- Percentage of tokens with correct head and label
- Different from standard Alpino evaluation scheme (CoNLL: single head per token)

Pre-Results (1): Sanity checks

- Evaluate parsers on CoNLL 2006 test data (368 sentences; brochure youth health; ASL 15.2)
- Using Alpino tags improved performance of data-driven parsers significantly
  (p < 0.002 according to Approximate Randomization Test with 1000 iterations)

Model | LAS | UAS | MRT (original CoNLL) | MRT (original CoNLL, cdb subpara) | MRT (original CoNLL, cdb subpara + Alpino)
--- | --- | --- | --- | --- | ---
Alpino | 78.05 | 82.92 | 78.57 | 82.71 | 79.14
Malt (correted w/ Alpino) | 82.45 | 80.51 | 80.64 | 82.66 | 81.79
Malt (correted w/ Alpino) | 79.23 | 80.51 | 79.59 | 80.11 | 78.59

Table: Performance of the data-driven parsers versus state-of-the-art performance.

Pre-Results (2): Trigram LM sentence perplexity

- Alpino (cor=0.069)
- MST (cor=0.87)
- Malt (cor=0.03)

Results (1): Average Sentence Length (ASL)

- All parsers robust to ASL, also grammar-based parser Alpino
- Zhang & Wang (2009): grammar-based parser (ERG) highly sensitive to ASL - longer sentences lead to sharp drop in parsing coverage

Results (2): Unknown Word Rate (UWR)

- Data-driven parsers sensitive to UWR
- Not the case for grammar-based parser: very good unknown word heuristics
- Note: Alpino UWR vs. simple UWR (UWR) - 0.07 for Alpino

Results (3): Unknown POS trigram rate (UPTR)

- Contrary to Zhang & Wang (2009): all parsers rather robust against UPTR

Results (4): Trigram LM sentence perplexity

- Additional measure: trigram LM estimated on cdb corpus
- Parsers are most sensitive to this measure
- If we remove two more possible outliers: Alpino -0.12, MST -0.57 and Malt -0.34

Discussion

- We added one measure, perplexity: shows highest correlation to parsing performance
  - Confirmed with much larger UWR (including Twenties Newspaper corpus, 500 million words): correlation with perplexity Alpino -0.23, MST -0.42 and Malt -0.47
  - Confirmed on DPC corpus (Dutch Parallel Corpus; various domains):

Conclusions & Future Work

- Measured correlation of parser performance with statistical measures of the text
- Simple measures, cheap to acquire
- Added perplexity measure: good indicator
- Could confirm general result found by Zhang & Wang (2009): Different parsing systems have different sensitivity against statistical measures of the text
- Started to look at 2nd question: Which system (grammar-based vs. data-driven) is more affected by domain shifts?

Future Work

- Simple measures quite predictive: Can we extend this line of work to identify subdomains?
- Domain detection
- More features: consider context (words, dependencies?) - unsupervised data clustering
- Data selection (related unlabeled data)
- Parse performance predictor as proxy for domain difference?