Improved statistical measures to assess natural language parser performance across domains

Barbara Plank
Computational Linguistics, University of Groningen (RUG), The Netherlands

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, etc.)?
- First step towards: How sensitive is a given system to the domain, i.e. which system (hand-crafted versus data-driven) is more affected by domain shifts?

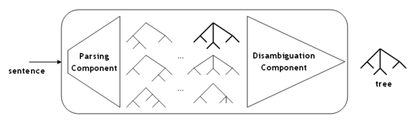
Related Work

- (Zhang & Wang, 2009): *Correlating Natural Language Parser Performance with Statistical Measures of the Text*. KI 2009.
  - How parser performance correlates to 3 simple statistical measures of the text
  - They looked at English (WSJ+Brown), examined several parsers (constituent, dependency, deep-grammar based)
- (Ravi, Knight & Soricut, 2008): *Automatic Prediction of Parsing Accuracy*. EMNLP 2008.
  - Build a regression model to predict parser accuracy (WSJ+Brown)

Parsing Systems

Alpino

- Hand-crafted grammar (HPSG-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 text 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:

Wikipedia	Wikipedia articles (excerpt)	#articles	#sentences	#words	ASL
LOC (location)	België, Brussel (stad)	31	2190	25259	11.5
KUN (arts)	School van Tervuren	11	998	17073	17.1
POL (politics)	Belgische verkiezingen 2003	16	983	15107	15.4
SPO (sports)	Spa-Francorchamps, Kim Clijsters	9	877	9713	11.1
HIS (history)	Geschiedenis van België	3	468	8396	17.9
BUS (business)	Algemeen Belgisch Vakverbond	9	405	4440	11.0
NOB (nobility)	Albert II van België	6	277	4179	15.1
COM (comics)	Suske en Wiske	3	380	4000	10.5
MUS (music)	Sandra Kim, Urbanus (artiest)	3	89	1296	14.6
HOL (holidays)	Feest Vlaamse Gemeenschap	4	43	524	12.2
Total		**95**	**6710**	**89987**	**13.4**

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. Marsi's software (CoNLL 2006): retag data with more fine-grained Alpino tags → positive effect
    Retagged data available at: `http://www.let.rug.nl/bplank/alpino2conll`
  - Simplifications (CoNLL): e.g. tokens have just a single head, MWUs 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):
- Percentage of unknown words, i.e. tokens not in cdb corpus (simple UWR, sUWR)
- For Alpino: percentage of words not in the lexicon (Alpino UWR, aUWR)

(3) Unknown Part-of-Speech trigram ratio (UPTR):
- Number of unknown PoS trigrams with respect to cdb corpus

(4) **Added**: Perplexity
- 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

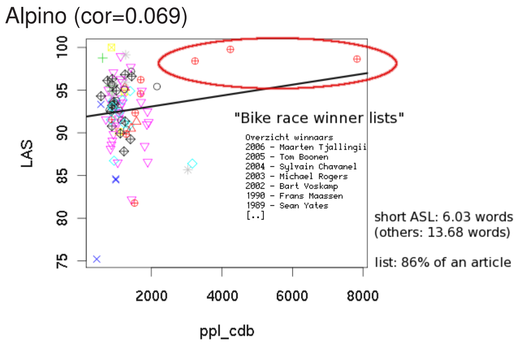
- Evaluate parsers on CoNLL 2006 test data (386 sentences; brochure youth health; ASL 15.2)
- Using Alpino tags improved performance of data-driven parsers significantly
  ($p < 0.002$ according to *Approximate Randomization Test* with 1000 iterations)

Model	LAS	UAS
MST (original CoNLL)	78.35	82.89
MST (original CoNLL, cdb subpart)	78.37	82.71
MST (cdb retagged with Alpino)	82.14	85.51
Malt (cdb retagged with Alpino)	80.64	82.66
MST (Nivre & McDonald, 2008)	79.19	83.6
Malt (Nivre & McDonald, 2008)	78.59	n/a
MST (cdb retagged with Mbt)	78.73	82.66

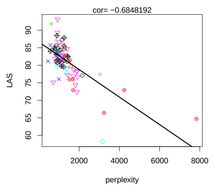
Table: Performance of the data-driven parsers versus state-of-the-art performance.

Pre-Results (2): Trigram LM sentence perplexity

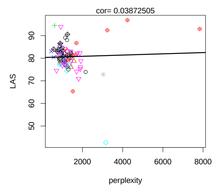
Alpino (cor=0.069)

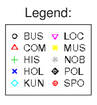


MST (cor=-0.67)



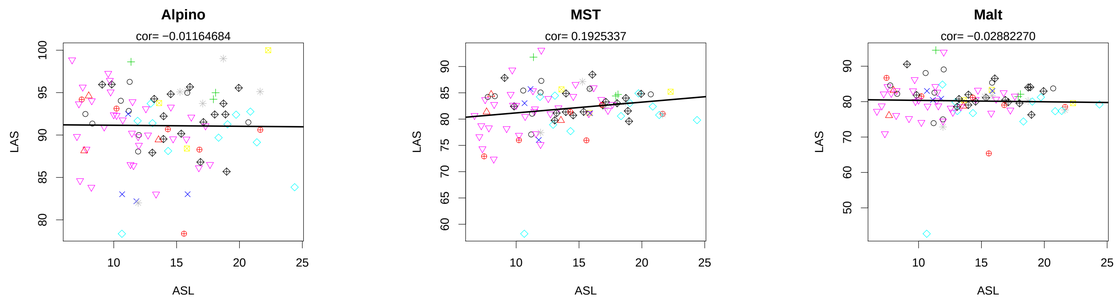
Malt (cor=0.03)





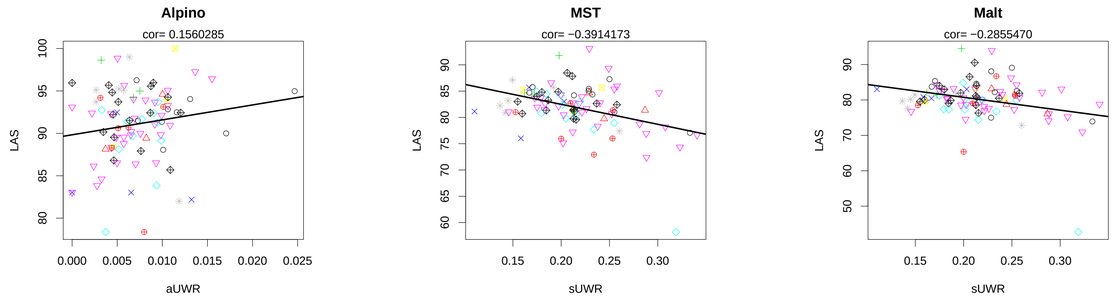
- 3 sport articles stand out
- For the remaining experiments, we excluded them

Results (1): Average Sentence Length (ASL)



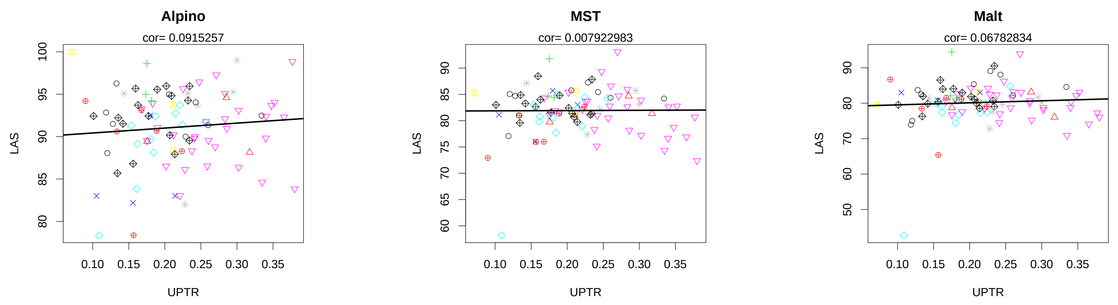
- All parsers robust to ASL, also grammar-based parser Alpino
- Zhang & Wang (2009): grammar-based parser (ERG) highly sensitive to ASL - longer sentences lead to sharp drop in parsing coverage

Results (2): Unknown Word Rate (UWR)



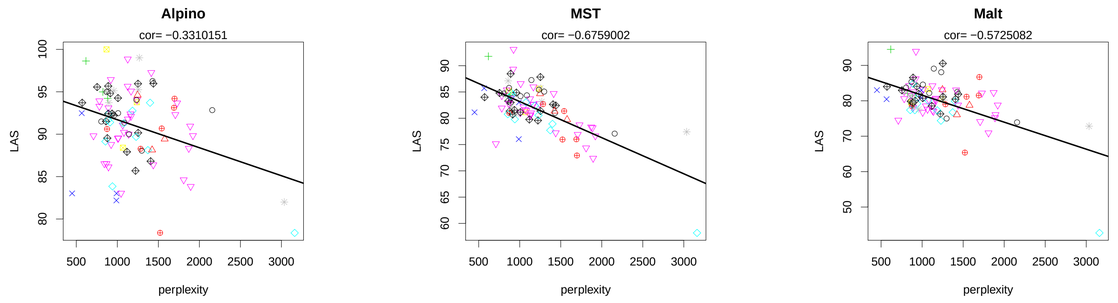
- Data-driven parsers sensitive to UWR
- Not the case for grammar-based parser: very good unknown word heuristics
- Note: Alpino UWR vs. simple UWR (sUWR) - sUWR: -0.07 for Alpino

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



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

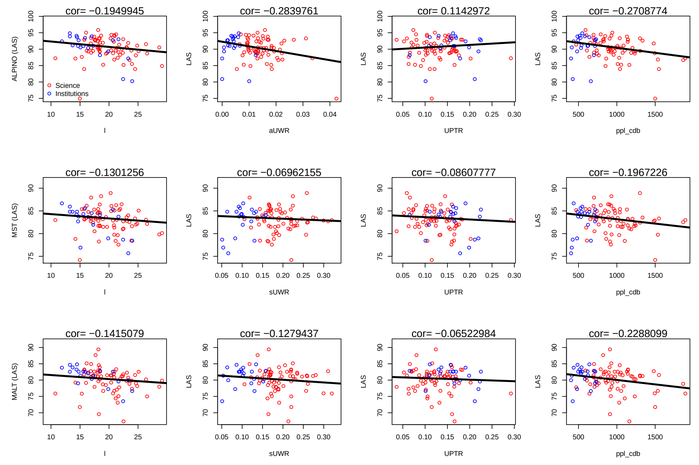
Results (4): Trigram LM sentence perplexity



- 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 LM (including Twente 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 content (words, dependencies?) - unsupervised data clustering
  - Data selection (related unlabeled data)
- Parse performance predictor as proxy for domain difference?

University of Groningen (RUG), The Netherlands E-Mail: b.plank@rug.nl Homepage: http://www.let.rug.nl/bplank