Lexical Association Analysis for Semantic-Class Enhanced Parsing

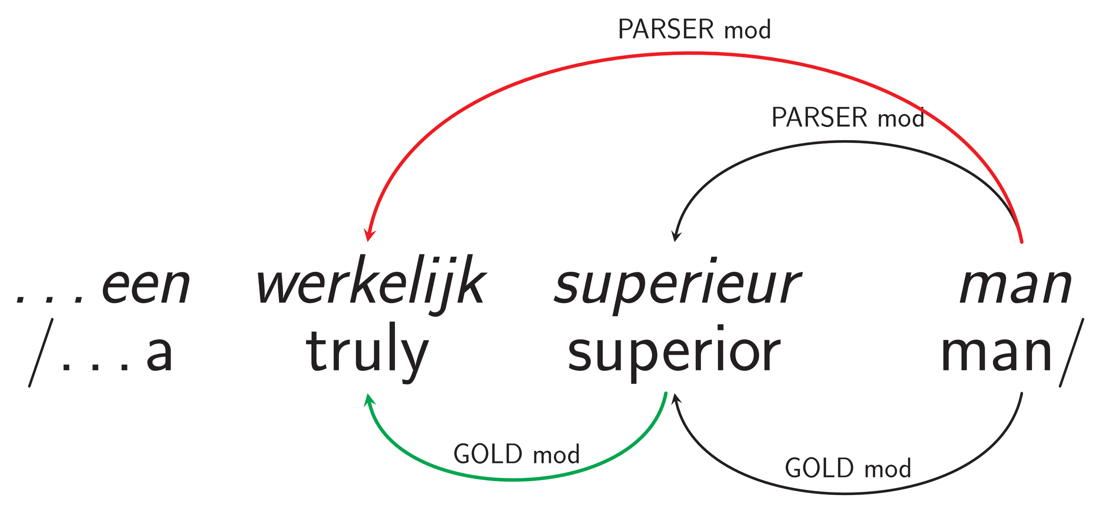
Simon Šuster and Gertjan van Noord
Language Technology, University of Groningen, The Netherlands

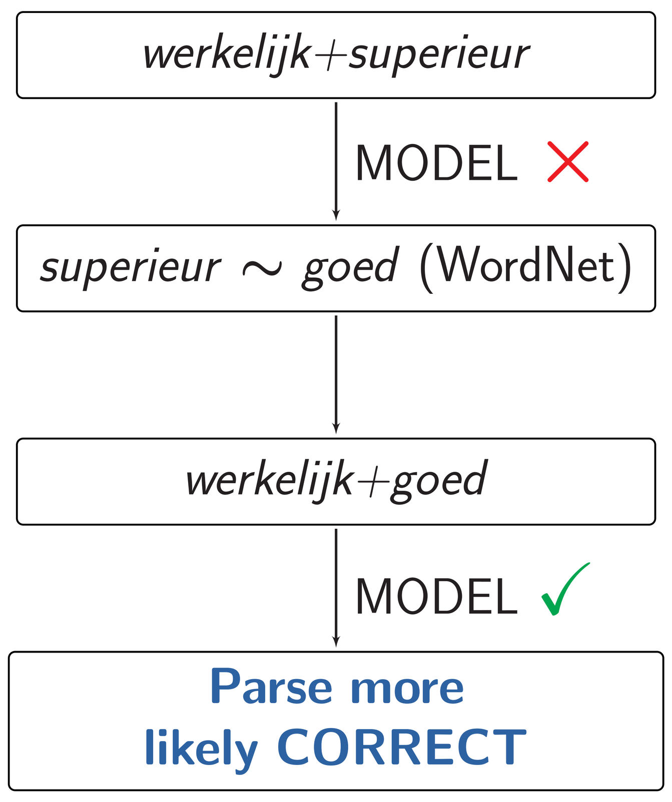
university of groningen

Introduction and Motivation

- Task: Parsing Natural Language
- Problem:

Association strength useful in parse disambiguation
But: Often missing due to lexical sparseness





- Goal: Identify *dependency types* with most space for improvement after enhancing the lexical-association model

Research Questions

- Coverage of 35 dependency types used in Alpino's lexical association model
- Which dependency types would contribute the most towards improved parsing accuracy with introduction of semantic classes)?
- Types of errors in parsed text
  - ▷ Can they be resolved by semantic classes?

Related Work

- (Agirre et al., 2011), (Henestroza & Candito, 2012), (Koo et al., 2008), (Candito & Seddah, 2010)

Parser

Alpino

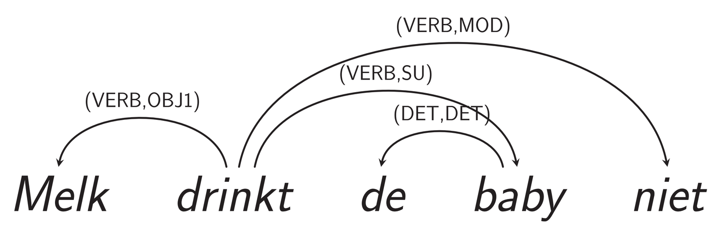
- Manually designed HPSG-like grammar for Dutch
- MaxEnt parse selection
  - ▷ One component is the lexical association model

Lexical association component ("the model")

- 35 dependency types of the form *(head-POS, dep-relation)*, holding between two words
- Examples:

(verb,obj1)	verb with a direct object
pp(noun,mod)	modification of a noun by a PP node
(adj,mod)	modification of an adjective

- Model provides MI-variant scores for instantiated dep. types



instance	dep. type	MI	weight
(drink,baby)	(verb,su)	.28	4.89
(drink,niet)	(verb,mod)	.16	3.02
(drink,melk)	(verb,obj1)	.39	4.66
(baby,de)	-	-	-

- Self-learning on a 500-million word corpus

Datasets

- Data for analysis - **Lassy Small**
  - ▷ 1-million words (from newspapers, Wikipedia, websites, fiction etc.), hand-annotated
- Model training data - **A 500-million word corpus**, parsed with Alpino

Statistical tests and evaluation

- *Coverage* (COV): proportion of test dependencies found in the model
- **Sentence-level analysis**: correlate COV_{sent} and *concept accuracy* (CA): non-parametric Spearman test
- **Dependency-level**: Pearson χ^2; Cramer's ϕ, odds ratio

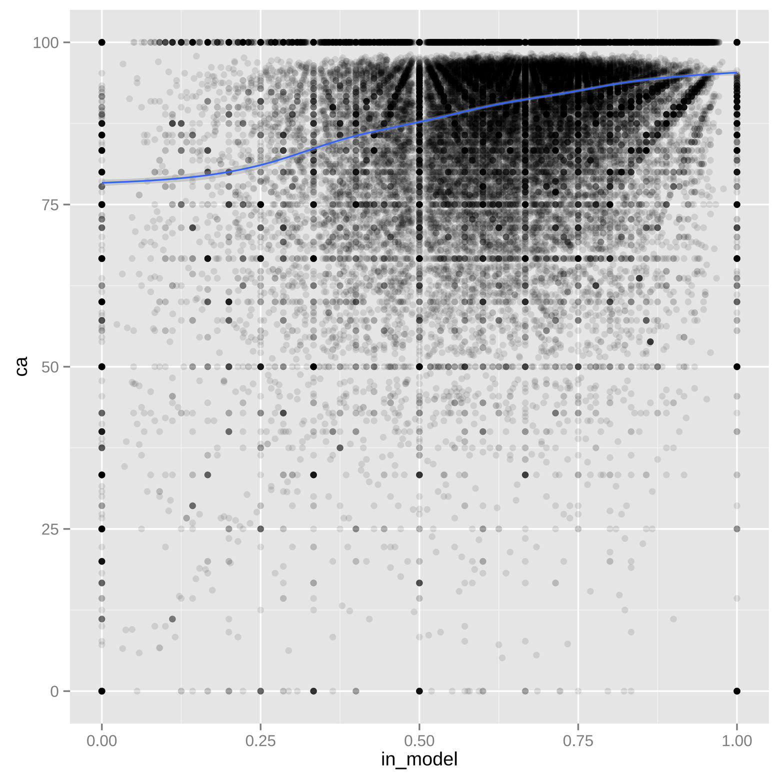
Coverage per Dependency Type

depType	# Lassy	# model	% model
(prep,obj1)	135645	115582	85.2
(noun,mod)	133925	86430	64.5
(verb,su)	113658	86640	76.2
(verb,mod)	96609	87591	90.7
(vg,cnj)	62658	54980	87.7
pp(noun,mod)	58561	17063	29.1
(verb,obj1)	49757	32586	65.5
(verb,vc)	41798	40877	97.8
pp(verb,mod)	41412	16261	39.3
...			
all	924783	672535	72.7

- Large variability: min. 17.5%, max. 97.8%
- Low % represents high lexical sparseness
- But: not all types are equally important in parse selection
- Include parsing performance

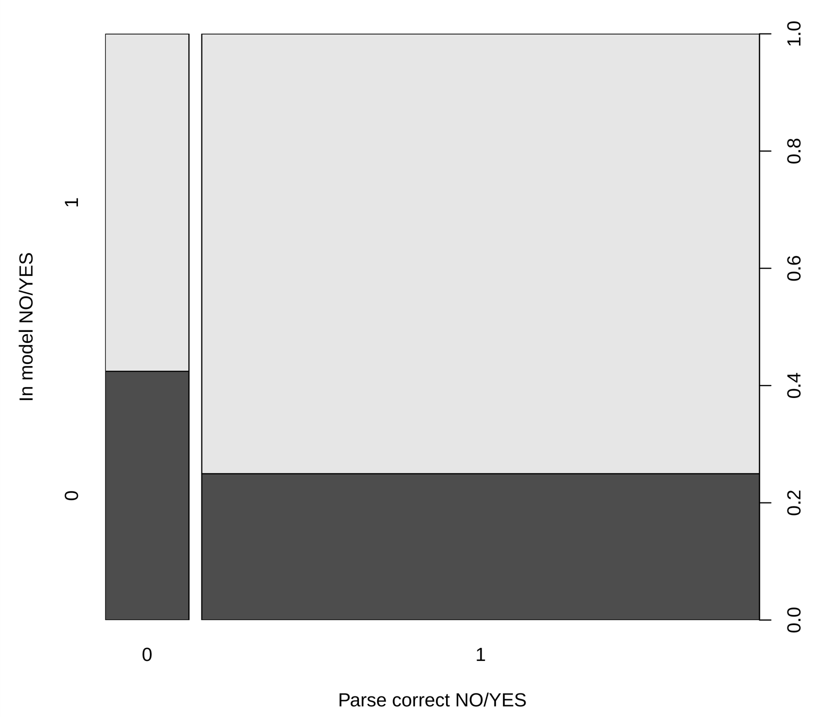
Results (1): Per-Sentence Perspective

- Mild correlation between COV and CA for all 62k sentences: $\rho_s = \mathbf{0.239}$
- Dotted-line effect due to analyzing proportions per sentence



Results (2): Per-Dependency View

- Association between parsing success and coverage for 925k dependencies



- χ^2 p $< \mathbf{0.001}$: parsing success on dependencies which *were* in the model is not the same as on the cases that *were not*
- Incorrect parse is **3.65 times more likely** when the dependency is not in the model (odds)

10 types with most space for improvement

depType	# deps	odds	ϕ coef.	χ^2-p	# 0-0	% 0-0
(verb,ld)	8079	14.962	.407	0	787	.097
(verb,pc)	14344	24.972	.368	0	583	.041
pp(verb,ld)	7468	2.925	.24	0	1636	.219
(adj,mod)	11828	2.653	.2	0	1213	.103
(noun,app)	11780	2.862	.193	0	2638	.224
(verb,predc)	20615	2.663	.141	0	719	.035
pp(verb,pc)	14795	1.829	.121	0	2058	.139
(noun,cnj)	18853	2.042	.12	0	3192	.169
(noun,mod)	133925	1.962	.11	0	8003	.06
(verb,obj1)	49758	1.776	.088	0	2431	.049

Selection criteria:

- Effect size and correlation should be relatively high
- # of incorrect parses not in model ("0-0" in table) should be high

Manual Verification

- Prevailing error type should be **wrong attachment**
- Many dependency types display mostly other error types
- Final set:
  - ▷ **(adj,mod)** – modification of the adjective
  - ▷ **(noun,cnj)** – coordination of nouns
  - ▷ **(noun,mod)** – modification of the noun

Conclusion

- **Nominal modification** and **coordination**, and **adjectival modification** most likely to aid the parser after their enhancement
- "Hard" attachment types: coordination of nouns shows up, but not PP attachment; no verbal types

Future work

- Develop a generalization method through distributional modeling and apply it to discovered dependency types

{S.Suster,G.J.M.van.Noord}@rug.nl