Daniel C. Müller

LEXICALIZED PARSING FOR DIFFERENT DOMAINS

10.12.2009

Laura Rimell and Stephen Clark

Hypothesis

"parser adaption

in the context of lexicalized grammar"

- according to two different domains

domains

- Biomedical domain
- Questions of question answering

Lexicalized parser

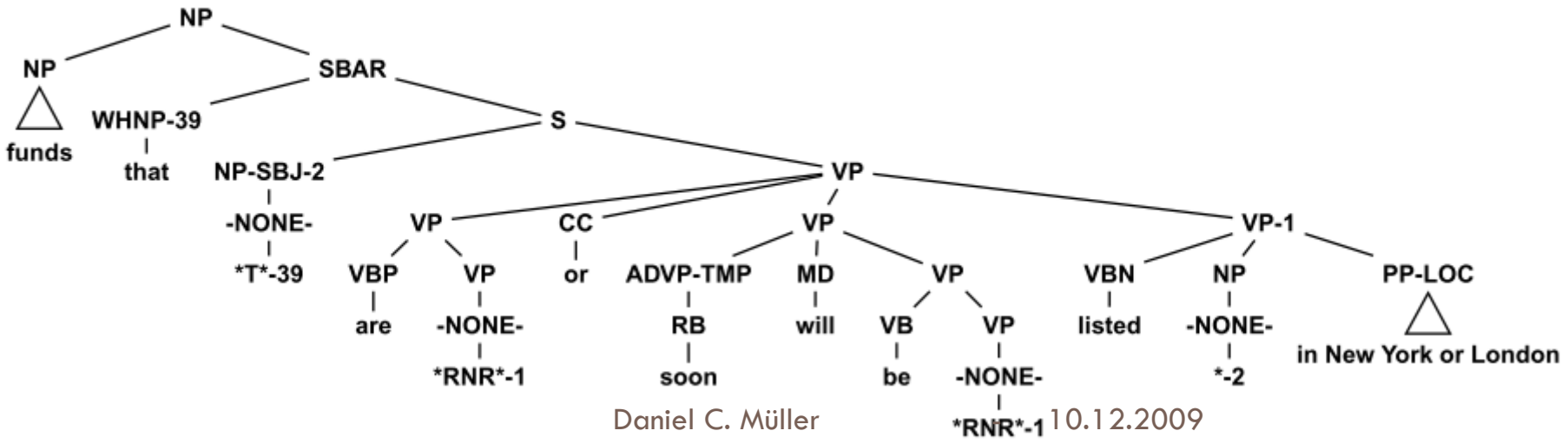
- POS-Tagging based on Penn Tree Bank
- Combinatory Categorial Grammar

  + manual annotation

Lexicalized parser

- POS-Tagging based on Penn Tree Bank
  - POS Tag:
    - 50 grammatical labels indicating part of speech
  - Each word → one POS Tag

funds that are or soon will be listed in New York or London.



Lexicalized parser

- POS-Tagging based on Penn Tree Bank
- Combinatory Categorial Grammar

  + manual annotation
  - lexical categorization (super-tagger)
    - 425 categories
    - Each word → at least one category
    - Containing subcategorial information
    - Complex categories like (S\NP)/NP means: returns S\NP when applied to NP

Lexicalized parser

- Example
  - Biomedical domain

POS Tag

Talin|NN	perhaps|RB	acts|VBZ	as|IN	a|DT	linkage|NN	protein|NN	.|.
NP	(S\NP)/(S\NP)	(S[dcl]\NP)/PP	PP/NP	NP[nb]/N	N/N	N	.

  - Question domain

What|WDT	king|NN	signed|VBD	the|DT	Magna|NNP	Carta|NNP	?|.
(S[wq]/(S[dcl]\NP))/N	N	(S[dcl]\NP)/NP	NP[nb]/N	N/N	N	.

lexical category

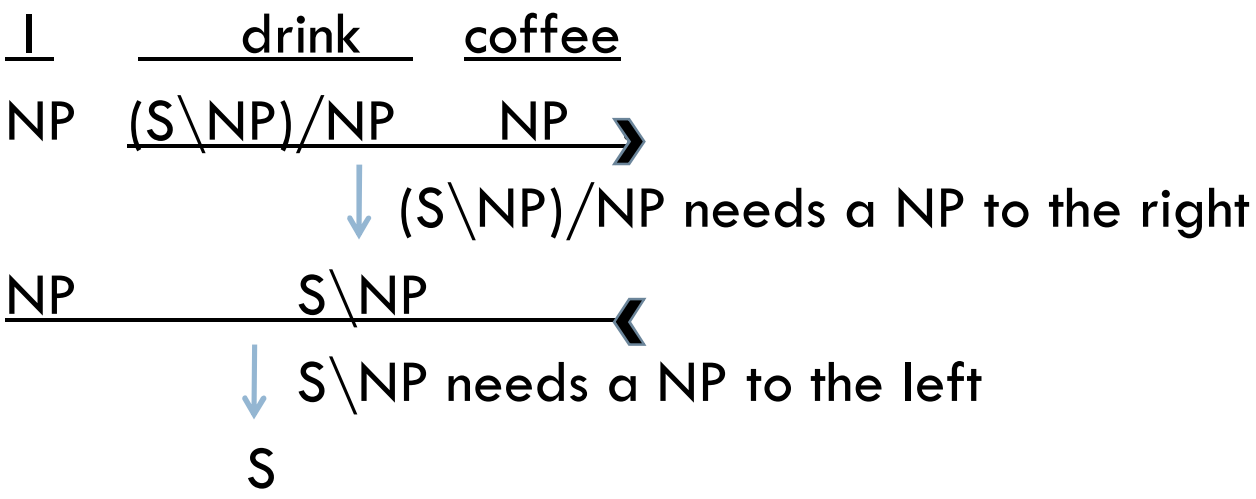
Lexicalized parser

- POS-Tagging based on Penn Tree Bank
- Combinatory Categorial Grammar

  + manual annotation
    - lexical categorization (super-tagger)
    - derivation (hierarchy)
      - Lexicalized categories + combinatory rules

Lexicalized parser

- Example



Motivation

- creating new training data

  at the lower levels of representation
  - better POS tagging → better categorization

- reduce annotation overhead

Experiments

- Training resources
  - Baseline
    - Wall Street Journal Sections 02-21 of CCGbank
  - Biomedical domain
    - POS tagger: gold-standard POS tags from GENIA
    - Lexical categories: first1,000 sentences of GENIA
    - parser evaluation: BioInfer
    - Evaluation set: Pyysalo et al. (2007b)
  - Question domain
    - Questions beginning with the word What, from the TREC 9-12 competitions: manually POS tagged & annotated with lexical categories

Experiments

- Results
  - POS-Tagger

%	WSJ 02-21	Retrained
Sec.00	96.7	-
Biomedical	93.4	98.7
Question	92.2	97.1

Experiments

- Results
  - Supertagger

%	Original pipeline	Retrained POS	Retrained POS & Super
Sec.00	91.5	-	-
Biomedical	89.0	91.2	93.0
Question	71.6	74.0	92.1

Experiments

- Results
  - Parser evaluation

%	Original pipeline	new POS	new POS & Super
Biomedical	76.0	80.4	81.5
Question	64.4	69.4	86.6

Analysis

- Comparing to WSJ:
  - Biomedical domain:
    - \+ similar syntactic structure
    - – vocabulary & foreign words
    - – long noun phrases
  - Question domain:
    - \+ vocabulary
    - – words with different distribution of POS in source domain
    - – **different syntactic structure**

Analysis

- POS tagging
  - Biomedical domain:
    - nouns and adjectives (801 NN + 268 JJ errors)

      very long noun phrases and unknown words like

      "major histocompatibility complex class II molecules"
  - Question domain:
    - wh-determiners (129 errors)

      only one occurrence in WSJ 02-21

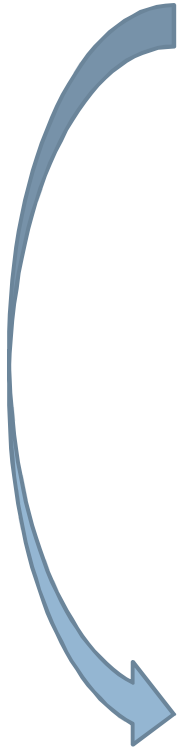
Analysis

- POS tagging
  - Biomedical domain:
    - nouns and adjectives (801 NN + 268 JJ errors)

      very long noun phrases and unknown words
  - Question domain:
    - wh-determiners (129 errors)

      (S/(S/NP))/N: "*What* Liverpool club spawned the Beatles?"

      S/(S\NP) : "*What* are the colors of the German flag?"



much more errors but related syntactic structure

Analysis

- Syntactic differences
  - Unknown POS n-gram rate

%	WJS 02-21		New training sets	
	3-grams	5-grams	3-grams	5-grams
Sec.00	0.4	12.1	-	-
Biomedical	0.7	10.9	0.5	9.2
Question	3.6	22.0	0.7	7.4

Analysis

- Syntactic differences
  - Unknown POS n-gram rate
  - Number of 20 most frequent POS n-grams

	3-grams	5-grams
Sec.00	18	19
Biomedical	16	13
Question	8	5

Analysis

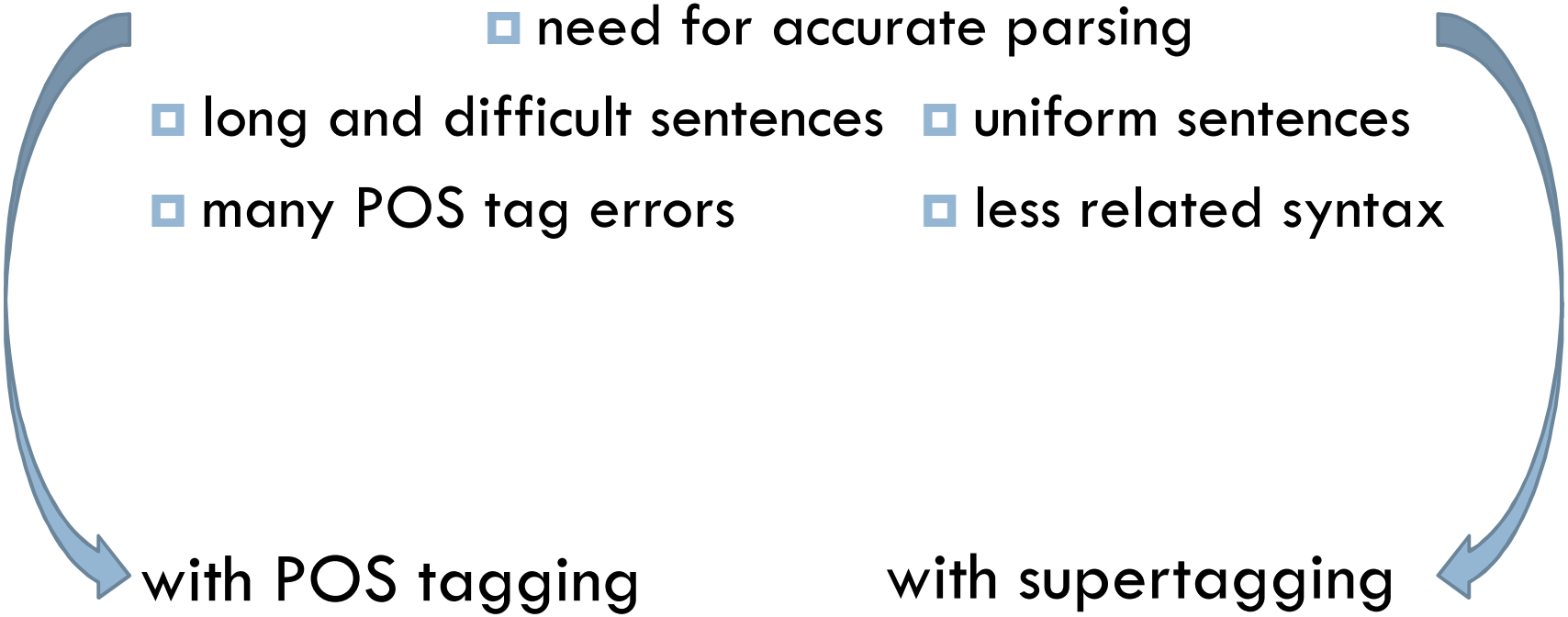
- Syntactic differences
  - Unknown POS n-gram rate
  - Number of 20 most frequent POS n-grams
  - POS Trigrams
    - Biomedical domain:
      - Domination of NPs and PPs
    - Question domain:
      - Beginning with WP VBZ like “What is”
      - Ending with VB . –

Analysis

- Syntactic differences
  - Unknown POS n-gram rate
  - Number of 20 most frequent POS n-grams
  - POS Trigrams
  - Number of rare or unseen lexical categories

Conclusion

- Biomedical domain
  - need for accurate parsing
  - long and difficult sentences
  - many POS tag errors
- Question domain
  - uniform sentences
  - less related syntax



with POS tagging

with supertagging

parser adaption successful!

References

- Laura Rimell, Stephen Clark. 2008. *Adapting a Lexicalized-Grammar Parser to Contrasting Domains. EMNLP 2008.*
- Julia Hockenmaiers. 2007. *Expressive Grammar Formalisms for Natural Language: Theory and Applications. Lecture 16: Extracting a CCG from the Penn Treebank.*
- Julia Hockenmaiers. 2005 CCGBank User's Manual