

Dependency-based Evaluation of Syntactic Parsers

Yusuke Miyao
University of Tokyo

Joint work with Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii

Background (1/2)

- Diversity of frameworks for syntactic parsing
 - Phrase structure parsers
 - Reranking parser [Charniak & Johnson 2005]
 - Lexicalized PCFG parser [Charniak 2000; Collins 1997]
 - State-splitting parser [Petrov & Klein 2007]
 - Dependency parsers
 - MST parser [McDonald & Pereira 2006]
 - Shift-reduce parser [Nivre 2005; Sagae & Tsujii 2007]
 - Deep linguistic parsers
 - LFG parser [Cahill et al. 2002; Kaplan et al. 2004]
 - CCG parser [Clark & Curran 2004]
 - HPSG parser [Oepen et al. 2004; Malouf & van Noord 2004; Miyao et al. 2004]

Background (2/2)

- Different evaluation criteria for different frameworks
 - Bracketing accuracy for phrase structure parsers
 - Attachment accuracy for dependency parsers
 - LFG parsers were evaluated with PARC 700 DepBank [King et al. 2003]
 - CCG parsers were evaluated with CCGBank [Hockenmaier et al. 2002]
 - HPSG parsers were evaluated with HPSG treebanks [Miyao et al. 2004; Oepen et al. 2004]
- Performances of these parsers differ?
- Framework-independent evaluation is possible?

Topic of this talk

- Comparative evaluation of Penn Treebank (PTB)-based phrase structure parsers and an HPSG parser
 - C&J: reranking parser [Charniak & Johnson 2005]
 - Charniak: lexicalized PCFG parser [Charniak 2000]
 - Enju: HPSG parser [Miyao et al. 2004]
- Approach: converting parser output into dependency-based syntactic representations

Parser evaluation schemes (1/2)

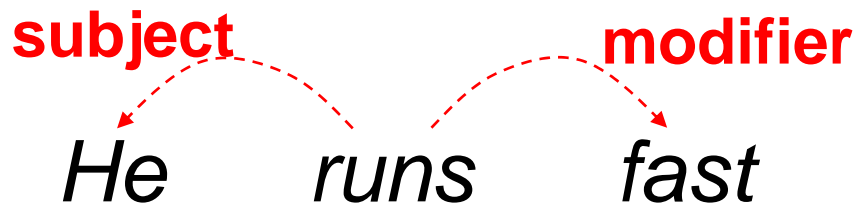
- Accuracy of phrase structures (brackets)
 - Most popular in the evaluation of PTB parsers
[Collins 1997; Charniak 2000; Charniak & Johnson 2005]
- Problems
 - Only surface structures are evaluated
 - Different grammar theories suppose different phrase structures

Parser evaluation schemes (2/2)

- Accuracy of labeled dependency relations
 - Closer to application needs
 - Deeper structures can be evaluated
- Standard in CoNLL shared tasks
- Popular in the evaluation of deep parsers, while using different gold-standard data [Kaplan et al. 2004; Burke et al. 2004; Clark & Curran 2004; Miyao et al. 2004]

Dependency-based evaluation

- We focus on two schemes of labeled dependency relations
 - Grammatical Relations (GR) [Briscoe & Carroll 2006]
 - Stanford Dependency (SD) [de Marneffe et al. 2006]
- Labeled dependencies among words



GR: example

- GR represents long-distance dependencies
- *They market cable-TV on the very grazing opportunities CNN seeks to discourage.*

...

(cmod _ opportunities seeks)

(ncsubj seeks CNN _)

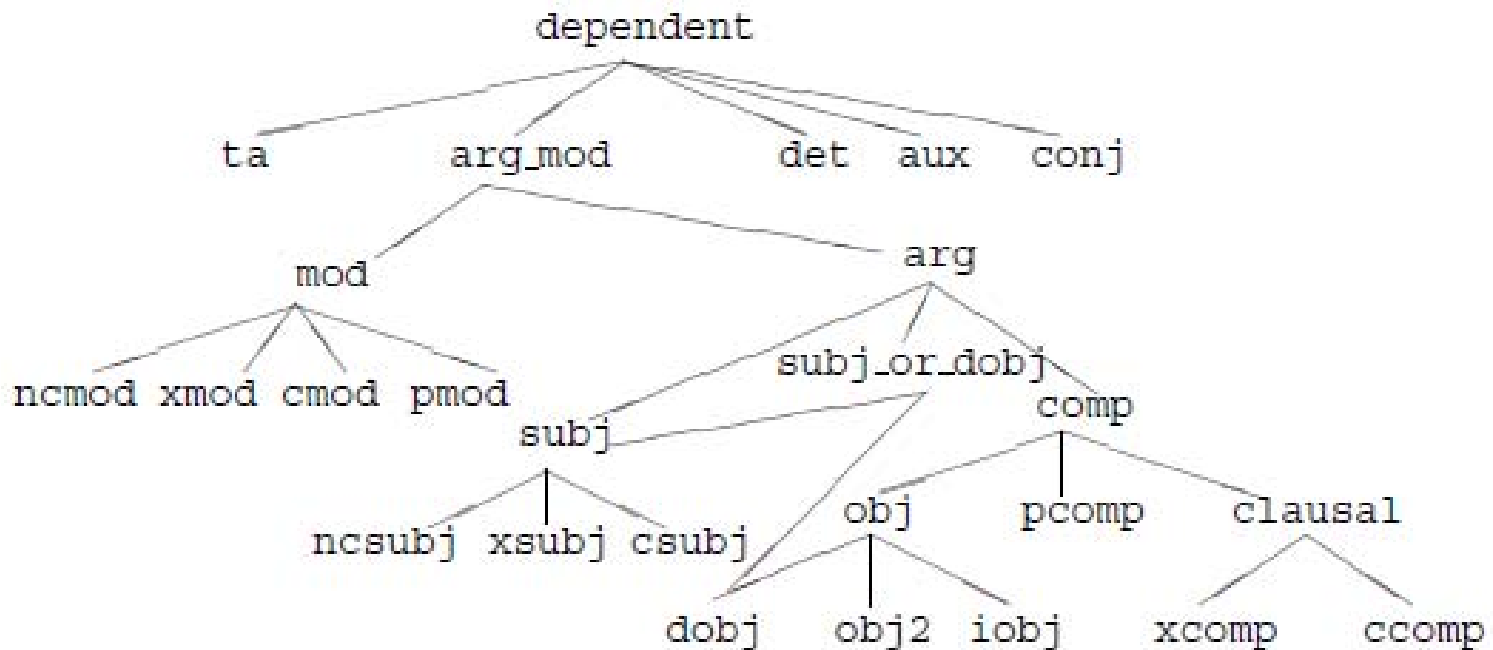
(ncsubj discourage CNN _)

(dobj discourage opportunities)

(xcomp to seeks discourage)

GR: hierarchy

- Relation types are organized in a hierarchy
- Non-leaf nodes allow for soft-matching of relation types



GR: evaluation metrics

- Microaverage precision/recall/f-score
 - Aggregate all relations including non-leaf types
 - Frequent relation types mainly affect the scores
- Macroaverage precision/recall/f-score
 - Average of accuracies for all relation types
 - Infrequent relation types can affect the scores

Stanford Dependency (SD)

- Originally designed for extracting useful relations from PTB-style phrase structures [de Marneffe et al. 2006]
- Provided as a program for converting PTB-style trees into labeled dependency relations
 - Attached to the Stanford parser [Klein & Manning 2003]
- Recently used for evaluating PTB parsers on biomedical text [Clegg et al. 2007; Pyysalo et al. 2007]

Example: *He runs fast*

nsubj(runs-2 He-1)
amod(runs-2 fast-3)

SD: example

- SD represents some long-distance dependencies, but not perfectly
- *They market cable-TV on the very grazing opportunities CNN seeks to discourage.*

...

```
nsubj(seeks-10, CNN-9)  
rmod(opportunities-8, seeks-10)  
aux(discourage-12, to-11)  
xcomp(seeks-10, discourage-12)
```

Example: GR and SD

Regulators also ordered CenTrust to stop buying back the preferred stock.

(ncsubj ordered Regulators _)
(ncsubj stop CenTrust _)
(ncsubj buying CenTrust _)
(ncmod _ ordered also)
(xcomp to ordered stop)
(xcomp _ stop buying)
(dobj buying stock)
(det stock the)
(passive preferred)
(ncsubj preferred stock obj)
(ncmod _ stock preferred)
(ncmod prt buying back)
(dobj ordered CenTrust)

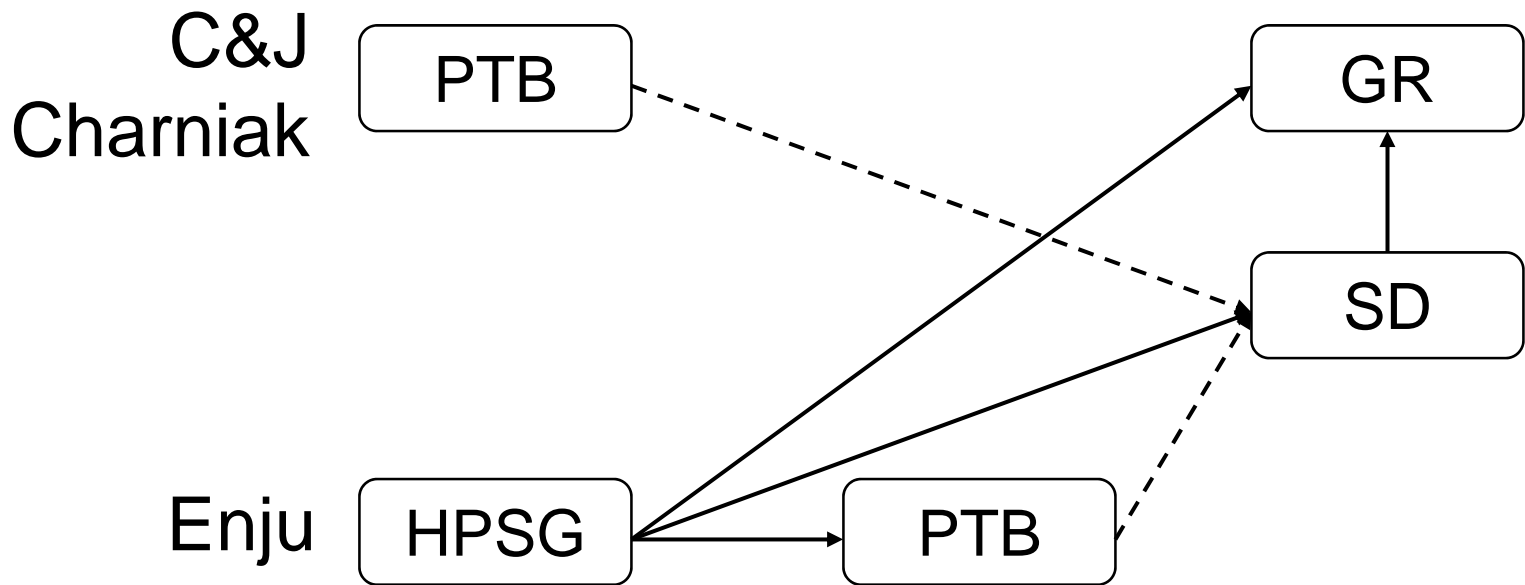
nsubj(ordered-3, Regulators-1)
advmod(ordered-3, also-2)
dobj(ordered-3, CenTrust-4)
aux(stop-6, to-5)
xcomp(ordered-3, stop-6)
partmod(stop-6, buying-7)
prt(buying-7, back-8)
det(stock-11, the-9)
amod(stock-11, preferred-10)
dobj(buying-7, stock-11)

Previous works on GR/SD evaluation

- RASP
 - Evaluated on GR [Briscoe & Carroll 2006]
- PTB parsers
 - Evaluated on GR (older version, Susanne) [Preiss 2003]
 - Evaluated on SD (GENIA and BioInfer) [Clegg et al. 2007; Pyysalo et al. 2007]
- CCG parser
 - Evaluated on GR [Clark & Curran 2007]
- LFG parsers
 - Evaluated on Parc 700 DepBank [Kaplan et al. 2004; Burke et al. 2004]

Our approach

- Format conversion: convert parser output into GR/SD, without changing the original parsers
 - SD-to-GR
 - HPSG-to-GR, HPSG-to-SD, HPSG-to-PTB



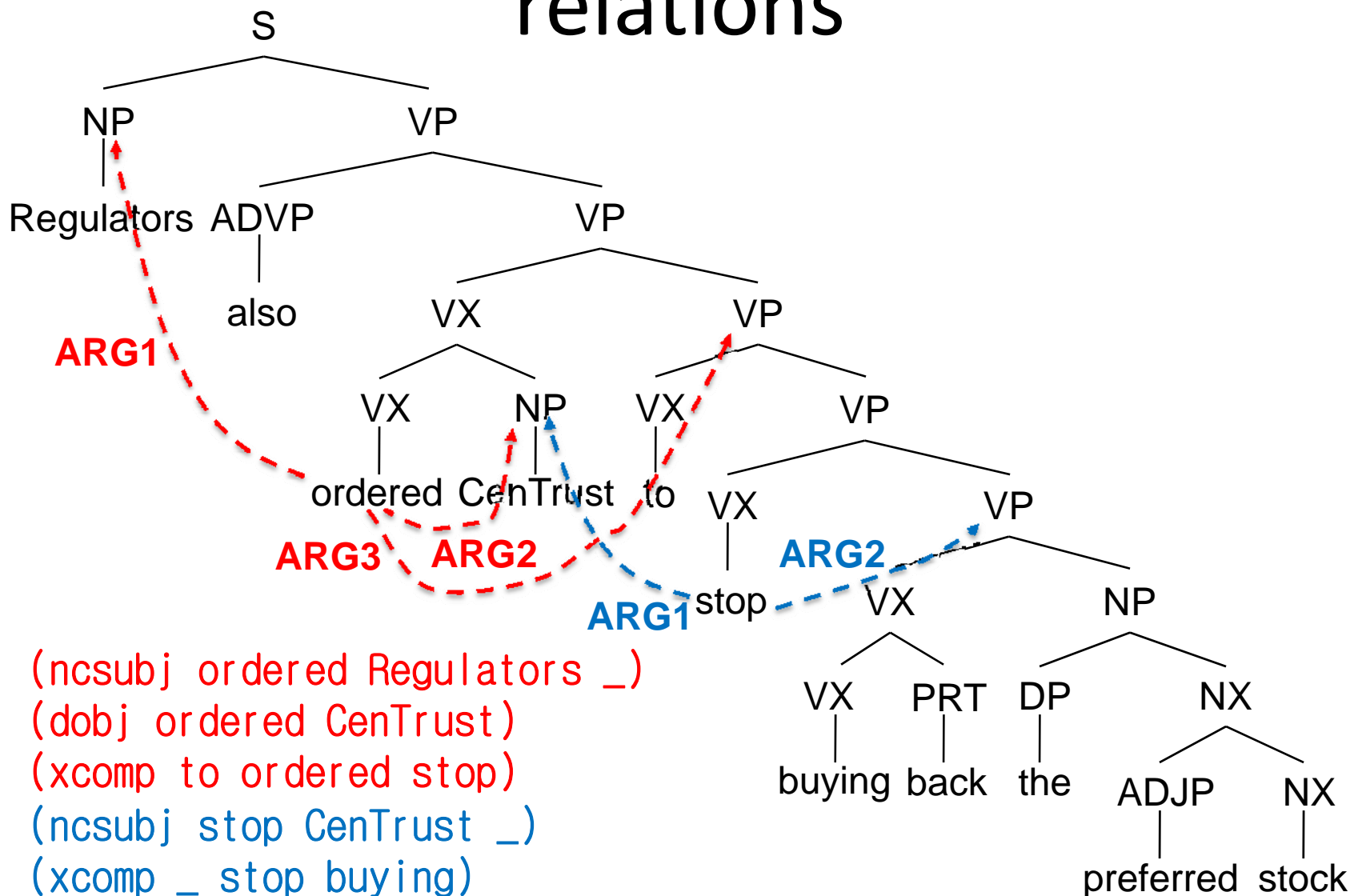
Our approach

- Use gold standard data (PTB and the HPSG treebank) when developing conversion programs
 - Accuracy obtained by converting a gold treebank indicates the quality of format conversion
- Conversion from SD to GR illuminates the difficulty in the translation between the two

HPSG-to-GR/SD conversion

- Basic strategy: mapping predicate-argument relations into GR/SD
- Before/after mapping, fix systematic disagreements of annotation policies
 - Add/remove dependency relations
 - Change lexical heads
 - Reduce coordinated relations
 - Construction-specific conversion

Mapping of predicate-argument relations



Add/remove dependency relations

- Add:
 - Relative-antecedent relations
 - Text adjuncts
- Remove:
 - Internal structure of named entities
 - Punctuations
 - Subjects of auxiliary verbs

Change lexical heads

- Some constructions needed changes of lexical heads

John came here ten years ago



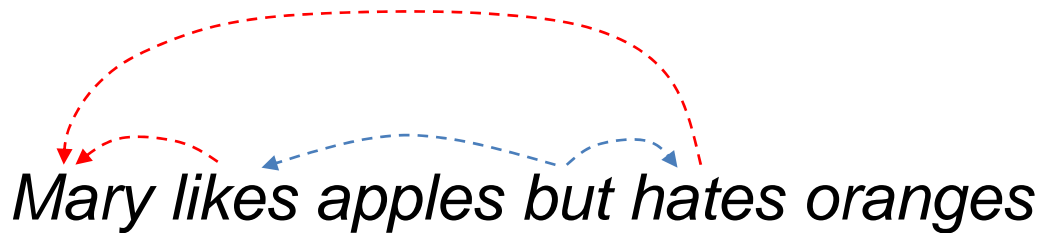
John came here ten years ago

- Heads of noun phrases (still remaining unfixed)

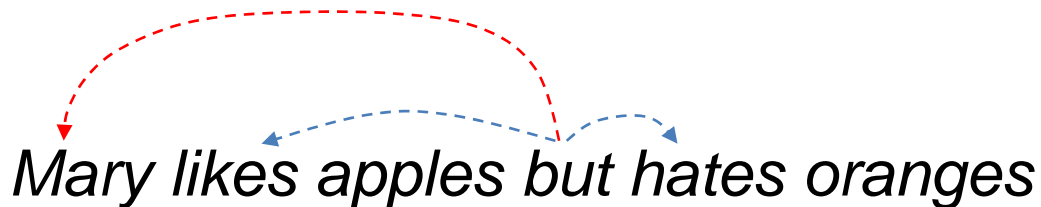
Reduce coordinated relations

- Shared arguments in coordinated phrases are reduced to one relation

Mary likes apples but hates oranges



Mary likes apples but hates oranges



Construction-specific conversion

- Quotation
- Copula
- Small clause
- Number expressions
- Relative clauses
- ...
- ~ 1,300 line Perl code

Experiments

- GR
 - 560 sentences from GR version of PARC 700 DepBank
 - Measure: microaverage, macroaverage
- SD
 - The same set of sentences as GR
 - Measure: accuracy
- Evaluated parsers
 - C&J [Charniak & Johnson 2005]
 - Charniak [Charniak 2000]
 - Enju [Miyao et al. 2004]
- We also show results for other parsers when available
 - RASP [Briscoe & Carroll 2006], C&C [Clark & Curran 2007], Stanford parser

Conversion accuracy

	Precision	Recall	F-score
SD→GR	80.84	69.16	74.54
HPSG→GR	87.49	86.79	87.14
CCG→GR [Clark & Curran 2007]	86.86	82.75	84.76

	Precision	Recall	F-score
HPSG→SD	83.43	81.44	82.42
PTB→SD (Stanford conversion)	100.00	100.00	100.00

	Precision	Recall	F-score
HPSG→PTB	98.12	98.07	98.09

GR evaluation

- Microaverage

	Precision	Recall	F-score
C&J (PTB→SD→GR)	79.08	67.46	72.81
Charniak (PTB→SD→GR)	78.41	67.68	72.65
Enju (HPSG→GR)	83.57	81.73	82.64
RASP [Briscoe & Carroll 2006]	77.66	74.98	76.29
C&C [Clark & Curran 2007]	82.44	81.28	81.86

- Macroaverage

	Precision	Recall	F-score
C&J (PTB→SD→GR)	60.20	47.97	53.39
Charniak (PTB→SD→GR)	59.39	48.08	53.14
Enju (HPSG→GR)	77.87	71.10	74.33
RASP [Briscoe & Carroll 2006]	62.12	63.77	62.94
C&C [Clark & Curran 2007]	65.61	63.28	64.43

SD evaluation

- Accuracy

	Precision	Recall	F-score
C&J (PTB→SD)	88.36	88.45	88.40
Charniak (PTB→SD)	87.05	87.10	87.07
Enju (HPSG→SD)	77.38	74.54	75.93
Enju (HPSG→PTB→SD)	87.13	87.16	87.14
Stanford parser	85.36	83.16	84.25

Format conversion errors

Comparison of
GR gold standard
and the converted
HPSG treebank

Remaining disagreements	107
text adjunct	35
argument/modifier distinction	34
lexical heads	25
POS	7
attachment ambiguity	6
Conversion errors	36
named entity	15
number expression	6
coordination	6
others	9
Errors of the HPSG treebank	14
noun phrase structure	10
others	4
Errors of GR data	13

Discussion

- Conversion quality >> parser performance
- Format conversion is non-trivial
 - Even conversion from SD to GR is difficult
 - Because of differences of lexical heads, POSs, etc.
- How to define annotation policies in a framework-independent way?
 - Without any framework, annotation decisions can be arbitrary
 - Un-documented decisions complicate the development of format conversion (arg/mod distinction, lexical heads, POS, etc.)
 - There are a lot of linguistically uninteresting constructions in real text

Conclusion

- Framework-independent dependency-based evaluation of syntactic parsers was described
 - Grammatical Relations
 - Stanford Dependency
- The deep parsers obtained higher scores in GR, while the shallow and deep parsers got similar scores in SD
- Accuracy figures are largely affected by the quality of format conversion
- Sources of difficulties in format conversion are presented
- Framework-independent evaluation is possible, but format conversion is much more critical

Future work

- Evaluation on other corpora
 - Susanne (GR data is available, although the annotation policy is slightly different)
 - Biomedical domain
- Better schemes for parser evaluation
 - For easier format conversion
 - Ambiguous/soft matching
 - Multiple reference annotations
 - For deeper representation
- Evaluation of dependency parsers
- Evaluating the impact of differences of parsing accuracy on applications