Institut für Computerlinguistik, Uni Zürich: Effiziente Analyse unbeschränkter Texte

Vorlesung 10: Evaluation

Gerold Schneider

Institute of Computational Linguistics, University of Zurich
Department of Linguistics, University of Geneva
gschneid@ifi.unizh.ch

December 15, 2003

Contents

1 Traditional Syntactic Evaluation: Labeled Bracketting

see Jurafsky & Martin 00: 464

PARSEVAL, Black et al. 1991

labeled recall: $\frac{\#\ of\ correct\ constituents\ in\ \textbf{candidate}}{\#\ of\ correct\ constituents\ in\ \textbf{gold standard}}$

labeled precision: $\frac{\#\ of\ \textbf{correct}\ constituents\ in\ candidate}{\#\ of\ \textbf{all}\ constituents\ in\ candidate}$

cross-brackets: *# of brackets overcrossing between candidate and gold standard*

2 Dependency-Based Evaluation: Lin 1995

PARSEVAL may count a single error multiple times:

a. [I [saw [[a man][with [[a dog] and [a cat]]]][in [the park]]]]
(let this be the gold standard)

b. [I [saw [[a man][with [[a dog] and [[a cat][in [the park]]]]]]]]
1 error: PP-attachment error to *cat* instead of *saw*, but 3 crossing brackets:

1. [a dog and a cat] vs. [a cat in the park]
2. [with a dog and a cat] vs. [a dog and a cat in the park]
3. [a man with a dog and a cat] vs. [with a dog and a cat in the park]

recall: 6/10. precision: 7/11.

c. [I [saw [a man] with [a dog] and [a cat][in [the park]]]]
very shallow, insufficient analysis, but no crossing brackets.
recall: 7/10. precision: 7/7.

Desiderata:

- Selective evaluation: depending on syntactic phenomena
- Ability to ignore inconsequential differences
- Facilitate the error diagnostics

$\longrightarrow$ Evaluation based on grammatical relations instead of constituency

3 An Annotation Scheme for Evaluation: Carroll et al. f.c.

3.1 More PARSEVAL problems

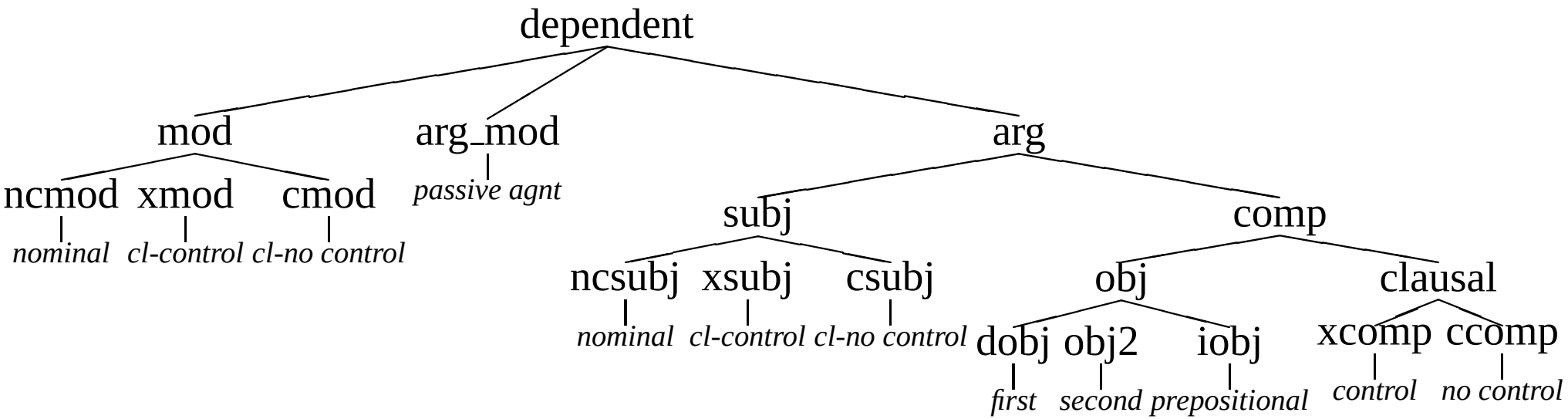
- Low agreement between parsing schemes for some constructions
  Partial PARSEVAL answer: remove certain bracketting info from consideration: negation, auxiliaries, punctuation, traces.
- Serious mapping problems to different annotation schemes remain

  “The treebanks have been constructed with reference to sets of informal guidelines indicating the type of structres to be assigned. In the absence of a formal grammar controlling or verifying the manual annotations, the number of different structural configurations tends to grow without check. For example, the [Penn treebank] implicitly contains more than 10000 distinct context-free productions, the majority occurring only once.”

- Penalises parsers that return more information than contained in the Treebank
- Cannot be applied to dependency-based parsers
- For cascading systems, different levels cannot be distinguished (chunking vs. parsing in my case).

3.2 Carroll et al. annotation hierarchy

(Carroll et al. f.c. 303, the $subj\_or\_dobj$ relation is left out)



cl: clausal

mod: modification, adjunct, arg: argument, complement

(no) control: *He*$_1$ *wants [*t_1 *to leave]* (control) vs. *He says [that she left]* (no control)

nc actually means non-clausal, but that mostly amounts to nominal incl. prepositional

The GRs are encoded as Lisp/Prolog facts. 500 random sentences from the Susanne Corpus.

Examples:

```
ncmod(_, flag, red).          % a red flag
ncmod(on, flag, roof).        % flag on the roof
xmod(without, eat, ask).      % he ate the cake without asking
cmod(because, eat, be).       % he ate the cake because he was hungry
arg_mod(by, kill, 'Brutus').  % killed by Brutus

ncsubj(she, eat, _).          % she was eating

xsubj(win, require, _).       % to win the America's Cup requires heaps of cash
csubj(leave, mean, _).        % that Nellie left meant she was angry
dobj(read, book,_).           % read books

dobj(mail, 'Mary', iobj).     % mail Mary the contract (3rd arg is initial_gr)
iobj(in, arrive, 'Spain').    % arrive in Spain
obj2(give,present,_).         % give Mary a present
xcomp(to, intend, leave).     % Paul intends to leave

xcomp(_, be, easy).           % Swimming is easy
xcomp(in, be, 'Paris').       % Mary is in Paris
ccomp(that, say, leave).      % I said that he left
```

4 First evaluation: tgrep-extraction-based

Using the grammatical relation (GR) data from the (held-out) section 00. Comparing the candidate parse GR and the tgrep'd GR. While the theoretical idea is fine, practical mapping problems occur:

- the tgrep patterns have (almost ?) 100 % precision, but below 100 % recall. (complexity problem)
- different grammatical assumptions (e.g. *in favour of, some of the people*)

The results reported are thus "about" 5 % too low.

5 Second evaluation: Mapping to Carroll et al.

Mapping to Carroll et al. is not always 1:1, but quite straightforward.

Naive direct mapping (c-subscript for Carroll relations):

$subj \Leftrightarrow ncsubj_c$, $obj \Leftrightarrow dobj_c$, $pobj \Leftrightarrow iobj_c$, $modpp \Leftrightarrow ncmod_c$ etc.

Works only partly:

- no adjunct/complement distinction for my PPs
- Tesnière translations
- different grammatical assumptions (e.g. Carroll does not consider relative pronouns to be subjects)

The mapping thus becomes more involved (as follows).

Mapping for subject and objects

Subject

Precision:	$subj$ OR $modpart$	$\rightarrow$	$ncsubj_C$ OR $cmod_C$(with rel.pro)
Recall:	$ncsubj_C$	$\rightarrow$	$subj$ OR $modpart$

$ncsubj_C$= non-clausal subject

$cmod_C$= clausal modification, used for relative clauses

Object

Precision:	obj OR $obj2$	$\rightarrow$	$dobj_C$ OR $obj2_C$
Recall:	$dobj_C$ OR $obj2_C$	$\rightarrow$	obj OR $obj2$

$dobj_C$=first object

$obj2_C$=second object

Mapping for PP-attachment

noun-PP

Precision:	$modpp$	$\rightarrow$	$ncmod_C$(with prep) OR $xmod_C$(with prep)
Recall:	$ncmod_C$(with prep) OR $xmod_C$(with prep)	$\rightarrow$	$modpp$

$ncmod_C$=non-clausal modification

$xmod_C$=clausal modification for verb-to-noun translations

verb-PP

Precision:	$pobj$	$\rightarrow$	$iobj_C$(with prep) OR arg_mod_C OR $ncmod_C$(with prep OR (prt & $dobj$)) OR $xcomp_C$(with prep) OR $xmod_C$(with prep)
Recall:	$iobj_C$(with prep) OR arg_mod_C	$\rightarrow$	$pobj$

$iobj_C$=prepositional object, arg_mod_C=passive agent

$xcomp_C$ for PP-attachment to copular verbs

6 Current Evaluation Results

Precision and recall measures		
subj_prec	828 of 946	87.5 %
subj_recall	767 of 956	80.2 %
obj_prec	430 of 490	87.7 %
obj_recall	316 of 391	80.8 %
nounpp_prec	343 of 479	71.6 %
verbpp_prec	350 of 482	72.6 %
ncmod_recall	593 of 801	74.0 %
iobj_recall	132 of 157	84.0 %
argmod_recall	30 of 41	73.1 %

Table 1: Pre-Current Evaluation of the Fully Lexicalized, Backed-Off System output

Current Evaluation and Comparison

	Percentage Values for			
	Subject	Object	noun-PP	verb-PP
Precision	91	89	73	74
Recall	81	83	67	83

Current selective Long-Distance Dependency (LDD) evaluation (as far as the annotations permit)

	LDD relations results for	
WH-Subject Precision	57/62	92 %
WH-Subject Recall	45/50	90 %
WH-Object Precision	6/10	60 %
WH-Object Recall	6/7	86 %
Anaphora of the rel. clause subject Precision	41/46	89 %
Anaphora of the rel. clause subject Recall	40/63	63 %
Passive subject Recall	132/160	83%
Precision for subject-control subjects	40/50	80%
Precision for object-control subjects	5/5	100%
Precision of $modpart$ relation	34/46	74%
Precision for topicalized verb-attached PPs	25/35	71%

7 Comparison to Related Work

7.1 Comparison to Lin

	Percentage Values for			
	Subject	Object	noun-PP	verb-PP
Precision	91	89	73	74
Recall	81	83	67	83
	Comparison to Lin (on the whole Susanne corpus)			
	Subject	Object	PP-attachment	
Precision	89	88	78	
Recall	78	72	72	

7.2 Comparison to Buchholz; and to Charniak (according to Preiss)

	Percentage Values for			
	Subject	Object	noun-PP	verb-PP
Precision	91	89	73	74
Recall	81	83	67	83
	Comparison to Buchholz ; and to Charniak, according to Preiss			
	Subject($ncsubj$)	Object($dobj$)		
Precision	86; 82	88; 84		
Recall	73; 70	77; 76		

7.3 Comparison to Carroll's Parser

only the numbers in bold can be compared

Relation	Precision	Recall
dependent	75	75
+mod	74	70
++ncmod	78	73
++xmod	70	52
++cmod	67	48
+**arg_mod**	**84**	**41**
+arg	77	84
++subj	84	88
+++**ncsubj**	**85**	**88**
+++xsubj	100	40
+++csubj	14	100
++comp	70	79
+++obj	68	79
++++**dobj**	**86**	**84**
++++obj2	39	84
++++iobj	42	65
+++clausal	73	78
++++xcomp	84	79
++++ccomp	72	75

8 Gradience. A Selection of Problematic Cases

Inter-annotator-agreement in the Carroll test corpus is “around 95 %”.

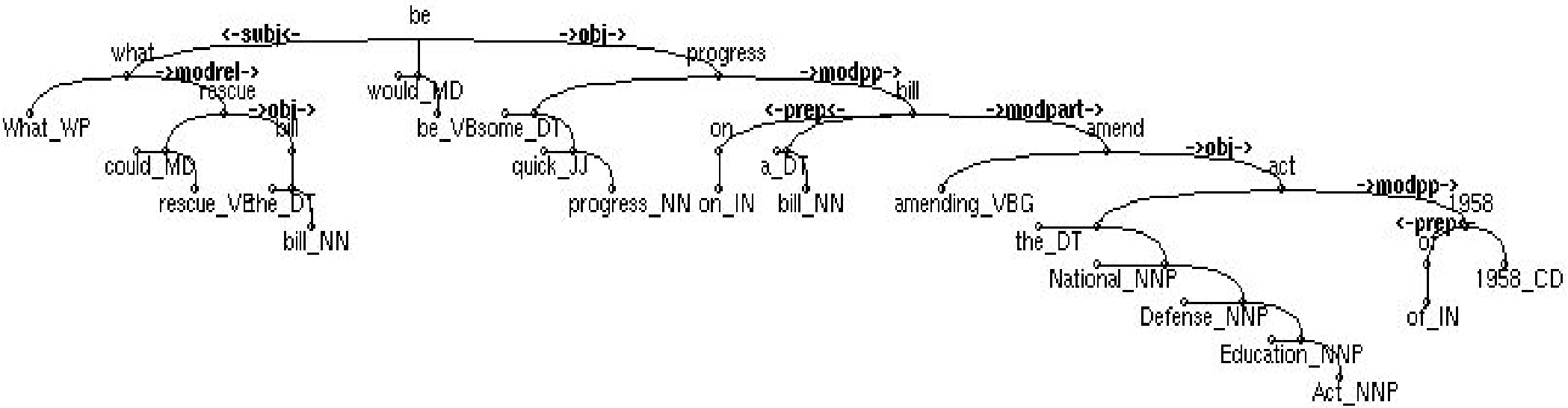


Figure 1: Aberrant but intentional analysis

csubj(’be’ , ’rescue’ , _ , 92). % gold standard: no ’what’ = ’that, which’ analysis

- *... the measure would provide means of enforcing the law ...*
  ncsubj(’enforce’ , ’measure’ , _ , 21).
  How far can control reach?
- *... there is nothing left of the conservative party ...*
  ncsubj(’nothing’ , ’leave’ , ’obj’ , 71). % gold standard: there-movement
  modpart(nothing, leave, _, ’→’, 71). % parser analysis: reduced relative
- *... prove [one of the difficult problems] ...*
  dobj(’prove’ , ’one’ , _ , 48). % gold standard: syntactical analysis
  obj(prove, problem, _, ’→’, 48). % parser analysis: hyperclever ‘semantic’ chunker, wrong head extraction?

- PP-attachment: ... *brought enthusiastic responses from the audience* ...
  ncmod('from' , 'response' , 'audience' , 11). % gold standard: to noun
  pobj(bring, audience, from, ('→'), 11). % parser: to verb
- PP-attachment: ... *(the government) made blunders in Cuba*
  ncmod('in' , 'blunder' , 'cuba' , 51). % gold standard: to noun
  pobj(make, cuba, in, '→', 51). % parser: to verb

  see discussion about PP-attachment in Lecture 3.