Part-of-speech tagging models for parsing

Rebecca Watson

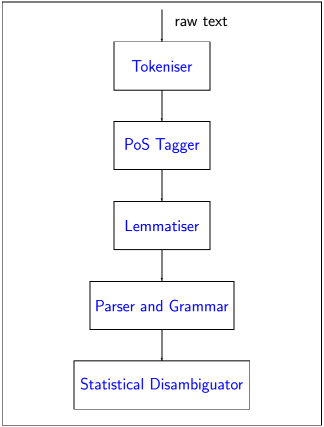
Computer Laboratory
University of Cambridge

March 7, 2006

Contents

1. RASP
2. Data
3. Experimentation
   - Tagger Experimentation
   - Parser Experimentation
4. Conclusions

Assuming a PoS tagger is used as a front end to parsing...



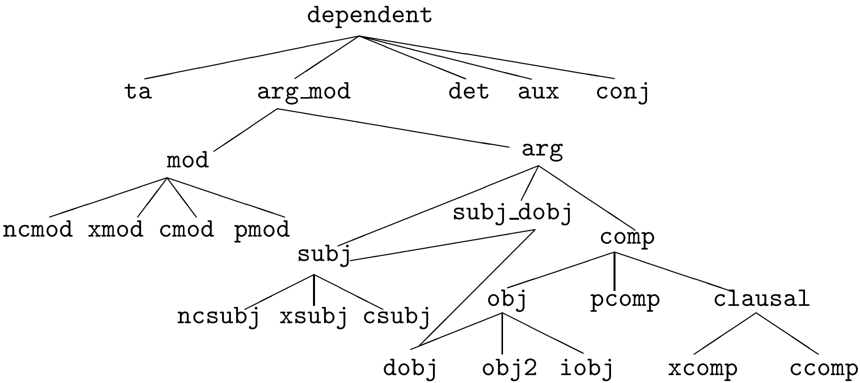
- Should the tag ambiguity be resolved by the tagger or be passed on to the parser?
- Can a parser perform PoS tagging more accurately than a PoS tagger?
- How does the chosen PoS tagging model affect parser performance?

Example

- Sentence Boundary Detection: We all walked up the hill.
- Tokenisation: We all walked up the hill .
- PoS Tagging
  - First order HMM PoS tagger using CLAWS II tagset: 149 PoS tags + 13 punctuation tags.
  - Single tag-per-word (tpw) output:
    We_PPIS2 all_DB2 walked_VVD up_RP the_AT hill_NN1 ._.
  - Multiple tpw output (with posterior tag probabilities):
    We(PPIS2:0.999983, NP1:1.73948e-05) all(DB2:0.803405, DB:0.168974, RR:0.0276206) walked(VVD:0.858121, VVN:0.141879) ...
  - Probability of a parse is the product of all shift/reduce action probabilities that resulted in creation of the parse *multiplied by the posterior tag probabilities.*
- Lemmatiser
  We_PPIS2 all_DB2 walk+ed_VVD up_RP the_AT hill_NNL1 ._.

Parser Output

- Syntactic tree → grammatical relations (GRs)
- Evalution scheme based on GRs.



Example

- We all walked up the hill.

(ncsubj walk+ed_VVD We_PPIS2 _)
(dobj walk+ed_VVD hill_NN1)
(det hill_NN1 the_AT)
(ncmod prt walk+ed_VVD up_RP)

- Single tpw (ncmod _ We_PPIS2 all_DB2)
- Multiple tpw (ncmod _ walk+ed_VVD all_RR)
- Parser selected correct PoS tag:
  all(DB2:0.803405, DB:0.168974, RR:0.0276206)
  The action probabilities out-weigh the tag probabilities.
  Therefore, worth passing on tag ambiguity if parser can perform PoS tagging more accurately than a PoS tagger.

Data

- Parc 700 Dependency Bank (DepBank): King *et al.* (2003)
- 560 sentence subset : outlined in Kaplan *et al.* (2004)
- DepBank560: Briscoe and Carroll (2005) extended DepBank with gold-standard GRs and (manually corrected) PoS tags.
- NE markup: for DepBank560 provided by Stephan Riezler, coauthor of Kaplan *et al.* (2004).

Gold Standards: We have both gold standard PoS tags and GRs so we can contrast PoS models both in terms of tagger and parser performance.

Contents

1. RASP
2. Data
3. Experimentation
   - Tagger Experimentation
   - Parser Experimentation
4. Conclusions

Tagging Experimentation

Can a parser perform PoS tagging more accurately than a PoS tagger?

This Work

- We will compare tagging accuracy of:
  - STAG and MTAG: PoS tagger with single or multiple tpw output.
  - MTAG-SYS and MTAG-SYS-DEF: Apply system thresholds over the MTAG input.
  - TOP-PARSE: top ranked parse (corresponding PoS tags).
  - NUM-TOP: PoS tags corresponding to the highest number of parses in the parse forest.
  - NUM-ALL: Normalised counts of NUM-TOP rank tags.
  - WEIGHT-TOP: Highest scoring tag based upon the (normalised) sum of probabilities of parses in which tags occur.
  - WEIGHT-ALL: Scores for WEIGHT-TOP rank tags.

Evaluation

- Standard measures: Precision and Recall.
- MRR: mean reciprocal rank of tags.
  $MRR = \frac{1}{\#tags} \sum_{i=1}^{\#tags} \frac{1}{correct-tag-rank_i}$
- Sent: The percentages of sentences containing at least one tagging error.

Results

- First four rows illustrate performance of the PoS tagger (also with RASP's thresholds).
- Upper bounds: provided by MTAG, errors here are caused by the unknown word module in the PoS tagger.

Tag Setup	Avg tpw†	Precision	Recall	MRR	Sent
STAG	1	97.23	97.23	97.18	40.71
MTAG-SYS-DEF (MSD)	1.12	88.50	98.79	97.94	21.79
MTAG-SYS (MS)	1.23	80.86	99.42	98.26	11.25
MTAG (M)	1.51	65.89	99.78	98.42	4.64
MSD-TOP-PARSE	1	95.38	95.38	95.38	59.11
MS-TOP-PARSE	1	94.47	94.47	94.41	64.46
M-TOP-PARSE	1	93.77	93.77	93.71	69.29
MSD-NUM-TOP	1	92.72	93.86	93.68	65.71
MSD-NUM-ALL	1.12	89.23	98.65	95.99	24.11
MSD-WEIGHT-TOP	1	94.67	95.84	95.66	54.82
MSD-WEIGHT-ALL	1.12	89.23	98.65	97.05	24.11

Table: Tagging Performance.†The average tag per word.

Results

- PoS tagger (STAG) outperforms all of the parser based PoS selection.
- Emulated performance over data with higher level of unseen words though same trends were witnessed.

Tag Setup	Avg tpw†	Precision	Recall	MRR	Sent
STAG	1	97.23	97.23	97.18	40.71
MTAG-SYS-DEF (MSD)	1.12	88.50	98.79	97.94	21.79
MTAG-SYS (MS)	1.23	80.86	99.42	98.26	11.25
MTAG (M)	1.51	65.89	99.78	98.42	4.64
MSD-TOP-PARSE	1	95.38	95.38	95.38	59.11
MS-TOP-PARSE	1	94.47	94.47	94.41	64.46
M-TOP-PARSE	1	93.77	93.77	93.71	69.29
MSD-NUM-TOP	1	92.72	93.86	93.68	65.71
MSD-NUM-ALL	1.12	89.23	98.65	95.99	24.11
MSD-WEIGHT-TOP	1	94.67	95.84	95.66	54.82
MSD-WEIGHT-ALL	1.12	89.23	98.65	97.05	24.11

Table: Tagging Performance.†The average tag per word.

Results

- Parser based models can't improve on the ranking of the PoS tagger either.

Tag Setup	Avg tpw[†]	Precision	Recall	MRR	Sent
STAG	1	97.23	97.23	97.18	40.71
MTAG-SYS-DEF (MSD)	1.12	88.50	98.79	97.94	21.79
MTAG-SYS (MS)	1.23	80.86	99.42	98.26	11.25
MTAG (M)	1.51	65.89	99.78	98.42	4.64
MSD-TOP-PARSE	1	95.38	95.38	95.38	59.11
MS-TOP-PARSE	1	94.47	94.47	94.41	64.46
M-TOP-PARSE	1	93.77	93.77	93.71	69.29
MSD-NUM-TOP	1	92.72	93.86	93.68	65.71
MSD-NUM-ALL	1.12	89.23	98.65	95.99	24.11
MSD-WEIGHT-TOP	1	94.67	95.84	95.66	54.82
MSD-WEIGHT-ALL	1.12	89.23	98.65	97.05	24.11

Table: Tagging Performance.[†]The average tag per word.

Contents

1. RASP
2. Data
3. Experimentation
   - Tagger Experimentation
   - Parser Experimentation
4. Conclusions

Parsing Experimentation

How does the chosen PoS tagging model affect parser performance (accuracy, coverage, efficiency)?

- Tagging experimentation illustrated that the parser could not perform PoS tagging as accurately (or efficiently) as the PoS tagger.
- However, tagging accuracy does not necessarily translate to equally detrimental parsing performance because the parser can recover from certain tag confusions and not others.

This Work

- Compare the parser's coverage, accuracy and efficiency given different PoS tag models.
- We will compare parser performance over several tagging models:
  - PoS tagger: STAG, MTAG, MTAG-SYS-DEF, MTAG-SYS.
  - Parser tagging models: NUM-TOP and WEIGHT-TOP (over MTAG-SYS-DEF).
- The impact of (Gold standard) PoS tagging and NE.
- Explore a hybrid (dynamic) tag selection model.

Evaluation

- Standard measures: Precision, Recall and F_1 (Accuracy).
- Frag: the proportion of sentences that result in a fragmentary parse (Coverage).
- Time: time taken to parse all 560 sentences (Efficiency).

Results

- Comparing PoS taggers:
  - Coverage vs. Efficiency vs. Accuracy
  - Best F_1 achieved by passing 1.12 tags per word (MTAG-SYS-DEF - tuned on Susanne).
  - Trade off between parse ambiguity and tag error rate.

Tag Setup	Prec	Rec	F_1	Frag	Time‡
STAG	71.06	70.96	71.01	21.25	0:03:50
MTAG-SYS-DEF	71.14	72.21	71.67	12.85	0:05:23
MTAG-SYS	70.10	71.39	70.74	10.00	0:18:27
MTAG	68.42	70.14	69.27	6.96	13:40:32
STAG-NE	73.53	69.66	71.54	25.00	0:03:13
MTAG-SYS-NE	72.54	70.49	71.50	12.68	0:10:57
MTAG-NE	71.32	69.30	70.30	9.28	0:45:51
MSD-WEIGHT-TOP	71.08	72.21	71.64	12.85	0:03:42
MSD-NUM-TOP	67.95	69.11	68.52	12.85	0:03:13
GOLD	72.94	73.12	73.03	14.46	0:04:39

Table: Parser Performance.‡Time as hours:minutes:seconds.

Results

- Compare performance of PoS taggers to parser based PoS tagging models:
  - PoS tagging models outperform parser based models in terms of accuracy and efficiency.

Tag Setup	Prec	Rec	F_1	Frag	Time[‡]
STAG	71.06	70.96	71.01	21.25	0:03:50
MTAG-SYS-DEF	71.14	72.21	71.67	12.85	0:05:23
MTAG-SYS	70.10	71.39	70.74	10.00	0:18:27
MTAG	68.42	70.14	69.27	6.96	13:40:32
STAG-NE	73.53	69.66	71.54	25.00	0:03:13
MTAG-SYS-NE	72.54	70.49	71.50	12.68	0:10:57
MTAG-NE	71.32	69.30	70.30	9.28	0:45:51
MSD-WEIGHT-TOP	71.08	72.21	71.64	12.85	0:03:42
MSD-NUM-TOP	67.95	69.11	68.52	12.85	0:03:13
GOLD	72.94	73.12	73.03	14.46	0:04:39
Upper Prec	82.25	31.34	45.39	-	4:02:49
Upper Rec	17.81	87.74	29.60		

Table: Parser Performance.‡Time as hours:minutes:seconds.

Results

- Compare impact of gold standard NE vs. gold standard PoS tagging.
  - Gain from PoS tagging far greater than that of NE recognition → effort should focus on improving PoS tagger. Though this may not be the case when there are a higher number of unseen words.

Tag Setup	Prec	Rec	F_1	Frag	Time‡
STAG	71.06	70.96	71.01	21.25	0:03:50
MTAG-SYS-DEF	71.14	72.21	71.67	12.85	0:05:23
MTAG-SYS	70.10	71.39	70.74	10.00	0:18:27
MTAG	68.42	70.14	69.27	6.96	13:40:32
STAG-NE	73.53	69.66	71.54	25.00	0:03:13
MTAG-SYS-NE	72.54	70.49	71.50	12.68	0:10:57
MTAG-NE	71.32	69.30	70.30	9.28	0:45:51
MSD-WEIGHT-TOP	71.08	72.21	71.64	12.85	0:03:42
MSD-NUM-TOP	67.95	69.11	68.52	12.85	0:03:13
GOLD	72.94	73.12	73.03	14.46	0:04:39
Upper Prec	82.25	31.34	45.39	-	4:02:49
Upper Rec	17.81	87.74	29.60		

Table: Parser Performance ‡Time as hours:minutes:seconds

Hybrid Selection

- Compare performance over non-fragmentary parses.

Tag Setup	Prec	Rec	F_1
STAG	73.66	74.94	74.30
MTAG-SYS-DEF	73.09	74.71	73.89
MTAG-SYS	72.07	73.49	72.77
MTAG	70.48	72.24	71.35
GOLD	74.58	75.70	75.14

Table: Performance over full parses.

- The increased performance illustrates that a large proportion of the errors are introduced by the frag parse output.
- The margin between STAG and GOLD has narrowed to only 0.84% F_1 suggesting that the tag errors account for a large proportion of the fragmentary parses.

Hybrid Selection

Can we rely on the grammar to find parses if and only if the correct tag sequence is input?

- Clark and Curran (2004):
  - apply a tag selection strategy where they assign a small number of supertags per word and increase the number of supertags if the parser fails to find an analysis.
  - increase efficiency, coverage and accuracy of the parser.

Hybrid Selection

- We combine the output from STAG (if full parses resulted) and MTAG-SYS-DEF (if fragmentary parses resulted for STAG).
- Increases accuracy and efficiency over MTAG-SYS-DEF (with same coverage).

Tag Setup	Prec	Rec	F_1	Frag[†]	Time[‡]
STAG	71.06	70.96	71.01	21.25	0:03:50
MTAG-SYS-DEF	71.14	72.21	71.67	12.85	0:05:23
MTAG-SYS	70.10	71.39	70.74	10.00	0:18:27
MTAG	68.42	70.14	69.27	6.96	13:40:32
GOLD	72.94	73.12	73.03	14.46	0:04:39
HYBRID	71.59	72.39	71.99	-	-

Table: Performance over full parses.

Conclusions & Future Work

- Conclusions:
  - Given a ‘good’ PoS tagger, parser-based tag selection models are unable to improve on the performance of the tagger or parser.
  - Multiple tpw input can increase parser accuracy and coverage but at a cost to efficiency.
  - Hybrid tag selection model provides a means to overcome the trade-off between tag error rates (coverage and accuracy) and increased parse ambiguity (efficiency and accuracy).
- Future Work:
  - Improve integration of the posterior tag probabilities with the parser’s statistical model.
  - Implement a dynamic model (extension of the hybrid model).

Acknowledgements

- Ted Briscoe and John Carroll.
- Funded by: Overseas Research Student Awards (ORS) and Poyton Scholarship (Cambridge Australia Trust)

TAGGING: Previous Work

- Charniak (1996):
  - 19 PoS tags (compared to 162 in CLAWS II)
  - Parser is only slightly more accurate than the tagger (96.1% vs. 95.9%).
- Dalrymple (2004):
  - Investigated the impact of PoS tags on parse ambiguity (number of parses).
  - Suggested that selecting the tag sequence corresponding to the largest number of parses may be the correct sequence (reducing ambiguity by around 50%).

PARSER: Previous Work

- Charniak (1996):
  - Parser is only slightly more accurate than the tagger (96.1% vs. 95.9%).
  - Parse Coverage: increases from 99.2% to 100% using multiple-tpw.
  - Efficiency: Four fold increase in the computational cost.