NLP

Introduction to NLP

The Penn Treebank

Description

- Background
  - From the early 90's
  - Developed at the University of Pennsylvania
  - (Marcus, Santorini, and Marcinkiewicz 1993)
- Size
  - 40,000 training sentences
  - 2400 test sentences
- Genre
  - Mostly Wall Street Journal news stories and some spoken conversations
- Importance
  - Helped launch modern automatic parsing methods

External Links

- Treebank-3
  - http://catalog.ldc.upenn.edu/LDC99T42
- Original version
  - http://catalog.ldc.upenn.edu/LDC95T7
- Tokenization guidelines
  - http://www.cis.upenn.edu/~treebank/tokenization.html
- The American National Corpus
  - http://www.americannationalcorpus.org/OANC/penn.html

Penn Treebank tagset (1/2)

Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d‘oeuvre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*'s*

Penn Treebank tagset (2/2)

Tag	Description	Example
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give *up*
TO	to	*to* go, *to* him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Example Sentence

- **`WSJ/12/WSJ_1273.MRG, sentence 11`**
- Because the CD had an effective yield of 13.4 % when it was issued in 1984 , and interest rates in general had declined sharply since then , part of the price Dr. Blumenfeld paid was a premium -- an additional amount on top of the CD 's base value plus accrued interest that represented the CD 's increased market value .

Parsed sentence

```
(S
  (SBAR-PRP
    (IN Because)
    (S
      (S
        (NP-SBJ (DT the) (NNP CD))
        (VP
          (VBD had)
          (NP
            (NP (DT an) (JJ effective) (NN yield))
            (PP (IN of) (NP (CD 13.4) (NN %))))
          (SBAR-TMP
            (WHADVP-4 (WRB when))
            (S
              (NP-SBJ-1 (PRP it))
              (VP
                (VBD was)
                (VP
                  (VBN issued)
                  (NP (-NONE- *-1))
                  (PP-TMP (IN in) (NP (CD 1984)))
                  (ADVP-TMP (-NONE- *T*-4)))))))))
                     ...
```

```
(S
  (SBAR-PRP
    (IN Because)
    (S
      (S
        (NP-SBJ (DT the) (NNP CD))
        (VP
          (VBD had)
          (NP
            (NP (DT an) (JJ effective) (NN yield))
            (PP (IN of) (NP (CD 13.4) (NN %))))
          (SBAR-TMP
            (WHADVP-4 (WRB when))
            (S
              (NP-SBJ-1 (PRP it))
              (VP
                (VBD was)
                (VP
                  (VBN issued)
                  (NP (-NONE- *-1))
                  (PP-TMP (IN in) (NP (CD 1984)))
                  (ADVP-TMP (-NONE- *T*-4))))))))
      (, ,)
      (CC and)
      (S
        (NP-SBJ
          (NP (NN interest) (NNS rates))
          (PP (IN in) (ADJP (JJ general))))
        (VP
          (VBD had)
          (VP
            (VBN declined)
            (ADVP-MNR (RB sharply))
            (PP-TMP (IN since) (NP (RB
then))))))))
```

```
(, ,)
(NP-SBJ
  (NP (NN part))
  (PP
    (IN of)
    (NP
      (NP (DT the) (NN price))
      (SBAR
        (WHNP-3 (-NONE- 0))
        (S
          (NP-SBJ (NNP Dr.) (NNP Blumenfeld))
          (VP (VBD paid) (NP (-NONE- *T*-3)))))))))
```

```
(VP
  (VBD was)
  (NP-PRD
    (NP (DT a) (NN premium))
    (: --)
    (NP
      (NP
        (NP (DT an) (JJ additional) (NN amount))
        (PP-LOC
          (IN on)
          (NP
            (NP (NN top))
            (PP
              (IN of)
              (NP
                (NP (DT the) (NNP CD) (POS 's))
                (NN base)
                (NN value))))))
      (CC plus)
      (NP (VBN accrued) (NN interest)))
    (SBAR
      (WHNP-2 (WDT that))
      (S
        (NP-SBJ (-NONE- *T*-2))
        (VP
          (VBD represented)
          (NP
            (NP (DT the) (NNP CD) (POS 's))
            (VBN increased)
            (NN market)
            (NN value)))))))
(. .))
```

```
(S
  (SBAR-PRP
    (IN Because)
    (S
      (S
        (NP-SBJ (DT the) (NNP CD))
        (VP
          (VBD had)
          (NP
            (NP (DT an) (JJ effective) (NN yield))
            (PP (IN of) (NP (CD 13.4) (NN %))))
          (SBAR-TMP
            (WHADVP-4 (WRB when))
            (S
              (NP-SBJ-1 (PRP it))
              (VP
                (VBD was)
                (VP
                  (VBN issued)
                  (NP (-NONE- *-1))
                  (PP-TMP (IN in) (NP (CD 1984)))
                  (ADVP-TMP (-NONE- *T*-4))))))))
      (, ,)
      (CC and)
      (S
        (NP-SBJ
          (NP (NN interest) (NNS rates))
          (PP (IN in) (ADJP (JJ general))))
        (VP
          (VBD had)
          (VP
            (VBN declined)
            (ADVP-MNR (RB sharply))
            (PP-TMP (IN since) (NP (RB
then))))))))
```

```
(, ,)
(NP-SBJ
  (NP (NN part))
  (PP
    (IN of)
    (NP
      (NP (DT the) (NN price))
      (SBAR
        (WHNP-3 (-NONE- 0))
        (S
          (NP-SBJ (NNP Dr.) (NNP Blumenfeld))
          (VP (VBD paid) (NP (-NONE- *T*-3)))))))))
```

```
(VP
  (VBD was)
  (NP-PRD
    (NP (DT a) (NN premium))
    (: --)
    (NP
      (NP
        (NP (DT an) (JJ additional) (NN amount))
        (PP-LOC
          (IN on)
          (NP
            (NP (NN top))
            (PP
              (IN of)
              (NP
                (NP (DT the) (NNP CD) (POS 's))
                (NN base)
                (NN value))))))
      (CC plus)
      (NP (VBN accrued) (NN interest)))
    (SBAR
      (WHNP-2 (WDT that))
      (S
        (NP-SBJ (-NONE- *T*-2))
        (VP
          (VBD represented)
          (NP
            (NP (DT the) (NNP CD) (POS 's))
            (VBN increased)
            (NN market)
            (NN value)))))))
(. .))
```

Peculiarities

- Complementizers
  - e.g., “that”
- Gaps
  - *NONE*
- SBAR
  - SBAR → COMP S
  - E.g., “that *NONE* represented the CD’ market value”

tgrep

`A < B`	A immediately dominates B
`A << B`	A dominates B
`A <- B`	B is the last child of A
`A <<, B`	B is a leftmost descendant of A
`A <<` B`	B is a rightmost descendant of A
`A . B`	A immediately precedes B
`A .. B`	A precedes B
`A $ B`	A and B are sisters
`A $. B`	A and B are sisters and A immediately precedes B
`A $.. B`	A and B are sisters and A precedes B

The Use Of Treebanks

- Disadvantages
  - A lot more work to annotate 40K+ sentences than to write a grammar.
- Advantages
  - Statistics about different constituents and phenomena
  - Training systems
  - Evaluating systems
  - Multilingual extensions

Introduction to NLP

Parsing evaluation

Evaluation Methodology (1/2)

- Classification tasks
  - Document retrieval
  - Part of speech tagging
  - Parsing
- Data split
  - Training
  - Dev-test
  - Test

Evaluation Methodology (2/2)

- Baselines
  - Dumb baseline
  - Intelligent baseline
  - Human performance (ceiling)
- New method
- Evaluation methods
  - Accuracy
  - Precision and Recall
- Multiple references
  - Interjudge agreement

Kappa

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- Agreement vs. expected agreement
  - P(A) is the level of agreement of the judges
  - P(E) is the expected probability of agreement by chance
- When $\kappa > .7$ – agreement is considered high
- Question
  - Judge agreement on a binary classification task is 60%, is this high?

Answer

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- Data
  - P(A) = .6
  - P(E) = .5
- Kappa
  - k = .1/.5 = .2
  - not high

Parsing Evaluation

- Precision and recall
  - get the proper constituents
- Labeled precision and recall
  - also get the correct non-terminal labels
- F1
  - harmonic mean of precision and recall
- Crossing brackets
  - (A (B C)) vs ((A B) C)
- PTB corpus
  - training 02-21, development 22, test 23

Evaluation Example

```
GOLD = (S (NP (DT The) (JJ Japanese) (JJ industrial) (NNS companies))
          (VP (MD should) (VP (VB know) (ADVP (JJR better)))) (. .)

CHAR = (S (NP (DT The) (JJ Japanese) (JJ industrial) (NNS companies))
           (VP (MD should) (VP (VB know)) ((ADVP (RBR better)))) (. .))
```

Bracketing Recall	=	80.00
Bracketing Precision	=	66.67
Bracketing FMeasure	=	72.73
Complete match	=	0.00
No crossing	=	100.00
Tagging accuracy	=	87.50

NLP