

Statistical Natural Language Parsing

Parsing: The rise of data and statistics



Pre 1990 (“Classical”) NLP Parsing

- Wrote symbolic grammar (CFG or often richer) and lexicon

$S \rightarrow NP VP$

$NN \rightarrow \textit{interest}$

$NP \rightarrow (DT) NN$

$NNS \rightarrow \textit{rates}$

$NP \rightarrow NN NNS$

$NNS \rightarrow \textit{raises}$

$NP \rightarrow NNP$

$VBP \rightarrow \textit{interest}$

$VP \rightarrow V NP$

$VBZ \rightarrow \textit{rates}$

- Used grammar/proof systems to prove parses from words
- This scaled very badly and didn't give coverage. For sentence:

Fed raises interest rates 0.5% in effort to control inflation

- Minimal grammar: 36 parses
- Simple 10 rule grammar: 592 parses
- Real-size broad-coverage grammar: millions of parses



Classical NLP Parsing: The problem and its solution

- Categorical constraints can be added to grammars to limit unlikely/weird parses for sentences
 - But the attempt make the grammars not robust
 - In traditional systems, commonly 30% of sentences in even an edited text would have *no* parse.
- A less constrained grammar can parse more sentences
 - But simple sentences end up with ever more parses with no way to choose between them
- We need mechanisms that allow us to find the most likely parse(s) for a sentence
 - Statistical parsing lets us work with very loose grammars that admit millions of parses for sentences but still quickly find the best parse(s)



The rise of annotated data: The Penn Treebank

[Marcus et al. 1993, *Computational Linguistics*]

```
( (S
  (NP-SBJ (DT The) (NN move))
  (VP (VBD followed)
    (NP
      (NP (DT a) (NN round))
      (PP (IN of)
        (NP
          (NP (JJ similar) (NNS increases))
          (PP (IN by)
            (NP (JJ other) (NNS lenders))))
          (PP (IN against)
            (NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
    (, ,)
    (S-ADV
      (NP-SBJ (-NONE- *))
      (VP (VBG reflecting)
        (NP
          (NP (DT a) (VBG continuing) (NN decline))
          (PP-LOC (IN in)
            (NP (DT that) (NN market))))))
      (. .)))
```



The rise of annotated data

- Starting off, building a treebank seems a lot slower and less useful than building a grammar
- But a treebank gives us many things
 - Reusability of the labor
 - Many parsers, POS taggers, etc.
 - Valuable resource for linguistics
 - Broad coverage
 - Frequencies and distributional information
 - A way to evaluate systems



Statistical parsing applications

Statistical parsers are now robust and widely used in larger NLP applications:

- High precision question answering [Pasca and Harabagiu SIGIR 2001]
- Improving biological named entity finding [Finkel et al. JNLPBA 2004]
- Syntactically based sentence compression [Lin and Wilbur 2007]
- Extracting opinions about products [Bloom et al. NAACL 2007]
- Improved interaction in computer games [Gorniak and Roy 2005]
- Helping linguists find data [Resnik et al. BLS 2005]
- Source sentence analysis for machine translation [Xu et al. 2009]
- Relation extraction systems [Fundel et al. *Bioinformatics* 2006]

Statistical Natural Language Parsing

Parsing: The rise of data and statistics