Statistical parsing in the third millennium: A brief history

*We'll take the millenium as beginning from 2000
I know that this is officially wrong

Christopher Manning

---


- There was nothing maximum entropy about it. It was a cleverly smoothed generative model
- Smooths estimates by smoothing conditional terms (which are a bit like maxent features):
  \[ P(t_t \mid e_{i..t}) \]
  \[ P(t_t \mid e_{i..t}, f_{p}) \]
- Biggest improvement is actually generative model predicts head tag first and then does \( P(w_t \mid \ldots) \).
- Gets 90.1% LP/LR F-score on sentences ≤ 40 wds

---


- Showed that one can successfully use a factored model to give effectively \( O(n^3) \) parsing of lexically-ized grammars by factoring phrase structure and lexical dependencies
- Showed how to do fast, guaranteed model optimal parsing of unlexicated and lexicated factored models using outside estimates and A* search
- Showed that lexicalization had been overemphasized while grammatical features and grammar refinement could give similar performance (86.3% LP LR F1)
- None of this actually sped up or improved parsers

---

Early discriminative parsing at Stanford

- Toutanova et al. (2002/2005) shows that a discriminative (maxent reranking) parser gives substantial gains over a generative parser in a stochastic HPSG grammar based parsing task
- Taskar et al. (2004 EMNLP) show how to do joint discriminative SVM-style ('max margin') parsing building a phrase structure tree also conditioned on words in \( O(n^3) \) time
- In practice, totally impractical and slow. Results were never demonstrated on sentences longer than 15 words

---

Charniak and Johnson (2005 ACL): Course-of-time Andrew parsing and MaxEnt discriminative reranking

- Builds a maxent discriminative reranker over parses produced by a (slightly bugfixed and improved version of Charniak (2000).
- Gets 50 best parses from Charniak (2000) parser
  - Doing this exploits the "course-of-time" idea to heuristically find good candidates
- Maxent model for reranking uses heads, etc. as generative model, but also nice linguistic features:
  - Contraction parallelism
  - Right branching preference
  - Heaviness (length) of constituents factored in
- Gets 91% LP/LR F1 (on all sentences!) - up to 80 wds

---

McDonald et al. (2005 ACL): Online Large-Margin Training of Dependency Parsers

- Builds a discriminative dependency parser
  - A different way to keep to \( O(n^3) \) time
  - Can again condition on rich features in that context
  - Again a true parser, not a n-best reranker
- Use of online large-margin training methods makes this way more practical than the Taskar et al. work
- Parser is faster than Charniak's!
- Doesn't report constituent LP/LR, but evaluating dependencies correct:
  - Accuracy is similar to but a fraction below Collins:
    - 90.9% vs. 91.4%