Uptraining for Accurate Deterministic Question Parsing

Slav Petrov
Google, New York

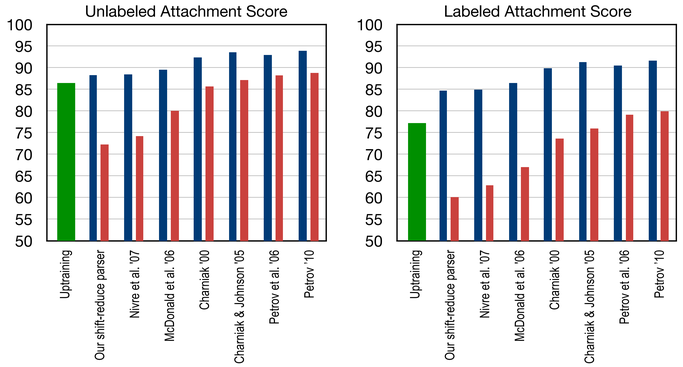
Pi-Chuan Chang
Google, Mountain View

Michael Ringgaard
Google, Aarhus

Hiyan Alshawi
Google, Mountain View

Motivation

We know that parsing performance goes down on out-of-domain text, but do you know *how* bad it can be?
It is very bad (at least for some parsers).

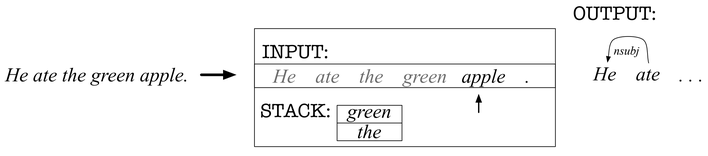


Various models trained on the Wall Street Journal and evaluated on the WSJ ▬ or on questions ▬.

Parsing Models

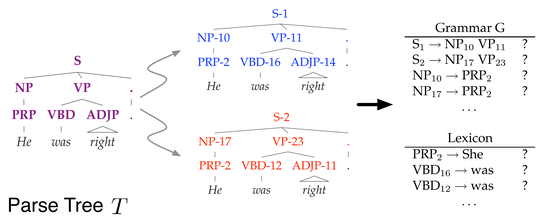
Deterministic Shift-Reduce Parser

In-house implementation of a shift-reduce parser (cf. Nivre's Malt Parser).



Latent Variable Parser

Refine the observed trees with latent variables and learn subcategories.



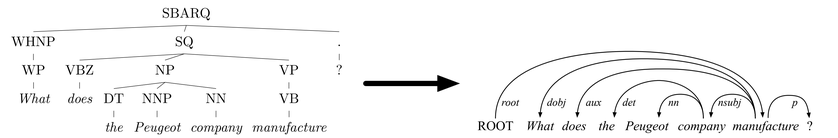
Parse Tree T
Sentence w

Derivations $t_i : T$

Model Parameters θ

Uptraining

Train a fast (but less accurate) parser on the output of a more accurate (but slower) parser.

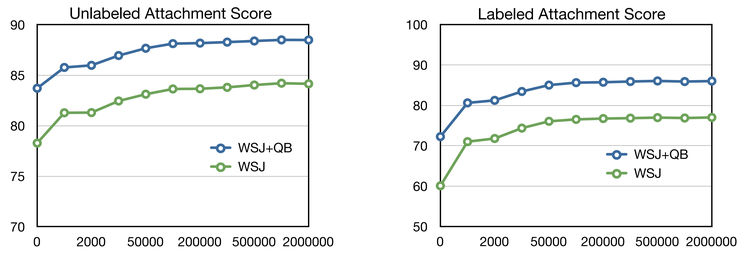


Parse a large amount of unlabeled text from the target domain with the Latent Variable Parser.

Convert the constituency trees to dependencies and retrain the shift reduce parser.

How Much Data?

- Wall Street Journal section of the Penn Treebank.
- QuestionBank (2K for training, 2K for evaluation).
- 1M questions from web queries that match a regular expresssion.
- Convert constituency trees to Stanford Dependencies.



Why Does It Work?

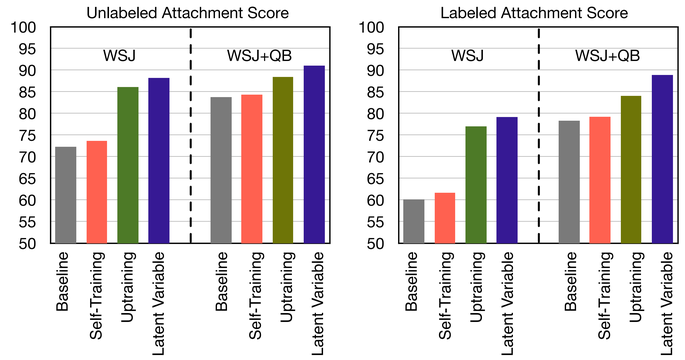
Hypothesis:
- Latent Variable Parser allocates some subcategories to questions and therefore generalizes better.

Two Experiments:
1. Collapse question specific phrasal categories (SQ, SBARQ) and retrain.
2. Remove all questions from the Wall Street Journal and retrain.

Results:
1. Latent Variable Parser loses 0.7% while Charniak's lexicalized parser loses 1.5%.
2. Charniak's lexicalized parser is better when no questions where present in training.

Final Results

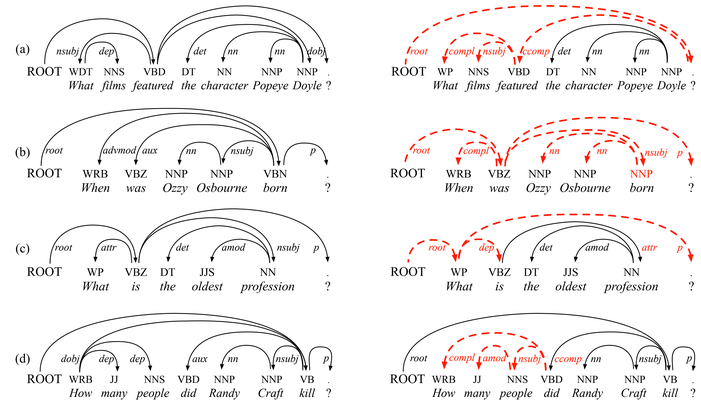
Uptraining with 100K unlabeled questions achieves comparable results to having 2K labeled questions. With 100K unlabeled and 2K labeled questions, uptraining closes the gap between in-domain and out-of-domain performance.



Various uptrained models evaluated on questions.

Examples

Example questions from the QuestionBank.
Gold parses on the left, predictions of a model trained on the WSJ on the right.
The *uptrained* model gets most of these right.



Google Research