On the statistical consintency of DOP estimators Prescher et. al.

Andreas van Cranenburgh (0440949)

Language & Optimality, University of Amsterdam

March 9, 2010

Abstract

Trouble in Data-Oriented Paradise

- The problem:
  - Current DOP estimators are inconsistent
  - An unbiased DOP estimator does not generalize
- However:
  - An unbiased estimator is futile
  - Achieving consistency will involve smoothing
- Results
  - Two new DOP estimators
  - Relation between DOP and memory-based models like k-nearest neighbor

Inconsistency

- Consistency: as the corpus size approaches infinity, increasingly likely to get arbitrarily close to the real distribution.
- definition:
  As corpus size goes to infinity, difference between estimated and actual probabilities approach zero.
- Bias: A deviation from a given probability model
- definition:
  Given a probability model, as corpus size goes to infinity, estimator's expectation equals the probability model.

DOP1 is both biased (prefers larger subtrees) and inconsistent (overfits).

Observations

- What could be a consistent estimator for DOP?
- Maximum-likelihood estimation is certainly consistent and unbiased,
- But: any unbiased DOP estimator does not generalize, ie., completely overfits
- This is contrary to the case with PCFGs, because DOP works with full parse trees instead of production rules.

Smoothing

- Other ways of achieving consistency:
- Select probability models using certain (a priori) parameters
  - But: contrary to DOP spirit, unknown what to select for
- Smoothing: reserve probability mass for unseen events
- Estimate probabilities of unseen events from seen events

Consistent estimators

Maximum-likelihood under resampling

1. Split treebank into held-out and extraction
2. Get subtrees from extraction part
3. Estimate their probabilities with MLE over held-out part
4. repeat steps 1-3 until convergence.

Problems with this approach:

- Only consistent if initial assignment is consistent (circular)
- Extremely inefficient

Consistent estimators

Backoff DOP

1. Start with full treebank, get relative frequencies
2. Discount them with eg., leave-one-out
3. Generate backoff treebank: all subtrees t1 and t2 that can be joined to form a tree from the treebank, assign probabilities from reserved prob. mass with backoff formula
4. Repeat steps 1-3 until trees in treebank can no longer be decomposed

Advantages of this approach:

- Combines likelihood with shortest derivation
- Trees in treebank are parsed as single look-ups, unseen trees using longer derivations with backoff
- Good empirical prospects