Ensemble-based Active Learning for Parse Selection

Jason Baldridge and Miles Osborne

School of Informatics, University of Edinburgh
{jbaldrid,miles}@inf.ed.ac.uk

December 15, 2003
Quick summary: 1

- **Active learning** is concerned with minimising the amount of annotated training material necessary to achieve a given performance level.

- With less training material:
  - We can create trainable speech and language technologies faster.
  - ... and save money.

- Labelling more training material will also lead to better results.
Quick summary: 2

Active learning results:

- Introduce multiple-model uncertainty sampling.
  - This easily outperforms (single-model) uncertainty sampling.

- Introduce a very simple active learning method – lowest best probability selection (LBP).
  - LBP is competitive with improved uncertainty sampling.
Quick summary: 3

Active learning results:

- Show that an ensemble trained without active learning can beat a single model trained with active learning.

- ... but that this ensemble can itself be outperformed by an ensemble trained with active learning.
Quick summary: 4

Parse selection results:

• For HPSG, an ensemble of three log-linear models achieves the best reported parse selection performance.

• Ad-hoc selection methods based upon superficial characteristics (sentence length, ambiguity rate etc) perform no better than random selection.

• Annotating sentences in the order they appear in the corpus is much worse than random selection.
Talk outline

• The English Resource Grammar (ERG) and the Redwoods Treebank.

• Parse selection for the ERG.

• Active learning (AL) methods.

• Experimental results.

• Comments
The English Resource Grammar

The ERG:

• ... is a broad-coverage manually written HPSG grammar.

• ... also provides semantic analyses of in-coverage sentences.
The Redwoods Treebank: 1

- Redwoods is a treebank of derivation trees for in-coverage sentences.

- Each such sentence has a distinguished preferred derivation tree.

- Derivation trees can be used to recover either parse trees or associated semantic interpretations.

- Latest version (3) statistics:

  | Sentences | Length | Parses |
  |------------|--------|--------|
  | 5302       | 9.3    | 58.0   |

- Only ambiguous sentences.
The Redwoods Treebank: 2

An example derivation tree
Talk outline

- The English Resource Grammar (ERG) and the Redwoods Treebank.
- Parse selection for the ERG.
- Active learning (AL) methods.
- Experimental results.
- Comments
A conditional log-linear model:

\[
P(t \mid s, M_k) = \frac{1}{Z(s)} \exp \left( \sum_{i=1}^{n} f_i w_i \right)
\]

Weights for model $M_k$ are determined using the LMVM algorithm (Malouf 02).

(We also use a perceptron model)
Parse selection: 2

- **Product model:**

\[ P(t \mid s, M_1, \ldots, M_n) = \frac{\prod_{i=1}^{n} P(t \mid s, M_i)}{Z} \]

- Based upon a **Product of Experts** formulation (Hinton 99).
  - . . . averages the contribution of each submodel.
  - . . . is an ensemble of log-linear models.
Parse selection: 3

- We treat the distribution of parses over a sentence in a binary manner.

- Three sets of features over derivations:
  - **Configurational**: loosely based on (Toutanova and Manning 02) – grandparent, local trees etc.
  - **Ngram**: derivations are flattened and treated as strings; ngrams are then extracted from these strings.
  - **Conglomerate**: features over phrase structure and Minimum Recursion Semantics (MRS).
Parse selection results

- Ten-fold cross-validation.
- Exact match evaluation.
- Unambiguous sentences are not counted.

| Method                | Score |
|-----------------------|-------|
| Random                | 22.7  |
| Log-linear (config)   | 74.9  |
| Log-linear (ngram)    | 74.0  |
| Log-linear (conglom)  | 74.0  |
| Product (all)         | 77.8  |
Talk outline

- The English Resource Grammar (ERG) and the Redwoods Treebank.
- Parse selection for the ERG.
- Active learning (AL) methods.
- Experimental results.
- Comments
Active learning

- The error of a model can be decomposed into a sum of:
  - **Noise**: intrinsic errors in the training set.
  - **Bias**: systematic errors a learner makes.
  - **Variance**: how much parameter estimates vary as a function of training set choice.

- Active learning methods generally select examples which reduce the variance of a model.
Active learning methods: 1

• Sample selection is one AL method.

• Basic idea:
  – Putatively automatically label all examples in a pool and select a subset of examples according to some method.
  – Manually label selected examples.
  – Remove labelled examples from the pool.
  – Retrain the model(s) and iterate.
Active learning methods: 2

- Sample selection for parse selection:
  - An example is a sentence.
  - Labelling an example means distinguishing one parse from the other parses for that sentence.

- *Annotation cost* is in terms of *selecting* the best parse (and not drawing parses from scratch).
Active learning methods: 3

- Selecting the best parse means navigating through a set of choice points.
- Each choice point (a discriminant) partitions the set of parses.
- A typical sentence requires 5 choices.
- Much more efficient than drawing a parse.
  - . . . implies that the best parse is present.
- Active learning annotation cost is in terms of the number of discriminants per sentence.
Uncertainty sampling: 1

- Tree entropy (Hwa 2000):

\[ f_{us}(s, \tau) = - \sum_{t \in \tau} p(t | s, M_i) \log p(t | s, M_i) \]

- Basic idea: selects examples with parses that are most uniformly distributed.

- Tree entropy has been applied to training CFG treebank parsers.

- We do not need to normalise tree entropy.
Uncertainty sampling: 2

- We can improve uncertainty sampling as follows:

\[ f_{us}^{es}(s, \tau) = -\sum_{t \in \tau} p(t | s, M_1, \ldots, M_n) \log p(t | s, M_1, \ldots, M_n) \]

- The single model has been replaced with a product (ensemble) model.

- We call this Product Uncertainty Sampling.
Lowest best probability selection

- LBP:

\[ f_{lbp}(s, \tau) = \max_{t \in \tau} p(t \mid s, M_i) \]

- Basic idea: selects examples with least discriminated parse.

- LBP is similar to uncertainty sampling.

- Generalising to an ensemble is trivial.
Query-by-committee

- Select examples when individual models predict different parses as being the preferred analysis.

- Basic idea: labelling uncertainly manifests as labelling disagreement.

- QBC is an ensemble method.
Talk outline

- The English Resource Grammar (ERG) and the Redwoods Treebank.
- Parse selection for the ERG.
- Active learning (AL) methods.
- Experimental results.
- Comments
Baselines

• For comparison we used the following baselines:
  
  – Select $n$ examples randomly.
  – . . . and label using a single model (config-random).
  – . . . and label using a product model (product-random).

• All experiments are averages over 10-fold cross-validation.

• Use $2k$ sentences.
Baseline results: 1

Random selection for a product model, Random selection for a single model
Baseline results: 2

- Random selection for our product model is better than random selection for a single model.

- Shows that improving the model can reduce annotation cost.
Main result: 1

US using a Π model, Random selection using a Π model, US using a single model
Main results: 2

- Random selection for our product model can outperform a single model with examples selected by active learning.

- . . . but ensemble-based active learning, for an ensemble model, outperforms random selection for an ensemble model.

- (A single model active learning method selecting examples for an ensemble model performs worse)
Heuristic selection

- Selecting shortest / longest / least ambiguous / most ambiguous sentences all performed no better than random selection.

- Selecting examples in the order they appeared in the corpus required 45% more labelling decisions than for random selection.
  - Most likely because Redwoods contains two domains.
## Cross method comparison: 1

| Method   | Cost | Reduction |
|----------|------|-----------|
| rand-config | 3700 | n/a       |
| rand-Π   | 1990 | 46.2%     |
| US-config | 2600 | 29.7%     |
| QBC      | 1300 | 64.9%     |
| LBP-Π    | 1280 | 65.4%     |
| US-Π     | 1300 | 64.9%     |

Annotation cost needed to achieve an average 70% parse selection performance.
## Cross method comparison: 2

| Method     | Cost | Reduction  | rand-config | rand-Π   |
|------------|------|------------|-------------|----------|
| rand-config| 13000| n/a        |             | (36.2%)  |
| rand-Π     | 8300 | 36.2%      | N/A         |          |
| US-config  | 7700 | 40.8%      | 7.2%        |          |
| QBC        | 3820 | 70.6%      | 54.0%       |          |
| LBP-Π      | 3660 | 71.9%      | 55.9%       |          |
| US-Π       | 3450 | 73.5%      | 58.4%       |          |

Annotation cost needed to achieve an average 75% parse selection performance.
Cross method comparison: 3

| Method     | Cost  | Reduction |
|------------|-------|-----------|
| rand-config | N/A   | N/A       |
| rand-$\Pi$  | 13800 | N/A       |
| US-config  | N/A   | N/A       |
| QBC        | 6780  | 50.9%     |
| LBP-$\Pi$  | 7320  | 47.0%     |
| US-$\Pi$   | 6410  | 53.6%     |

Annotation cost needed to achieve an average 77% parse selection performance.
Comments

- Active learning can dramatically reduce the annotation effort involved with training HPSG parse selection mechanisms.

- Ensemble methods can improve both parse selection and active learning.

- Further reductions should follow from only considering $n$-best parses.

- Ongoing work is concerned with bootstrapping a semantic interpretation system based on the ERG (Rosie Project).