How to Select a Good Training-data Subset for Transcription: Submodular Active Selection for Sequences — Lin & Bilmes

Partha Lal

20th October 2009/Listen!
Task description

- Given an unlabeled corpus $V$ and a budget to label $K$ utterances, choose the optimal set $S$ to label $S \subseteq V$.

Prior work

**Uncertainty Sampling** A small amount of data is labeled and a model trained. Further data is labeled based on the uncertainty of the model on those samples.

**Query-by-Committee** A set of models is created from a small amount of labeled data. The samples with the greatest disagreement between committee-members is then labeled by hand.
Submodularity

- Function $z$ is submodular if

$$ (S \cup \{s\}) - z(S) \leq (R \cup \{s\}) - z(R) $$

(1)

where $R \subseteq S \subseteq V$

- *Diminishing returns* — if adding an element to a set $S$ raises $z(S)$, then adding the same element to a superset of $S$ won’t raise it by more (roughly speaking)

- The discrete version of convexity
Facility location

\[ z_1(S) = \sum_{i \in V} \max_{j \in S} w_{i,j} \] (2)

Graph cut

\[ z_2(S) = \sum_{i \in V \setminus S} \sum_{j \in S} w_{i,j} \] (3)

- A greedy search algorithm is used to grow \( S \) until budget \( K \) is used up.
- If a submodular function is nondecreasing and normalized then the greedy algorithm will only be worse than optimal by some constant fraction, and will run in polynomial time.
Fisher Kernel

\[ w_{i,j} = \kappa(i, j) \]  \hspace{1cm} (4)

- Allows us to compare sequences of differing length
- Fisher score for utterance \( X_i \) is a vector \( U_i \) with length equal to the number of parameters \( \theta \)

\[ U_i = \frac{\partial}{\partial \theta} \log p(X_i | \theta) \]  \hspace{1cm} (5)

\[ w_{i,j} = - \| U_i - U_j \|_1 \]  \hspace{1cm} (6)
Results

- TIMIT phone recognition
- No language model used
- 16 component GMMs used
- The initial model used for Fisher scores was (a) unsupervised and then (b) supervised
Results without initial model

Figure 1: Relative improvements over the average phone error rate of random selection. No initial model scenario.
Results with initial model

Figure 2: Relative improvements over the average phone error rate of random selection. With initial model scenario.