Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning

Yo Ehara  
NICT  
ehara [at] nict.go.jp  
http://yoehara.com/

Yusuke Miyao  
NII

Hidekazu Oiwa  
Issei Sato  
Hiroshi Nakagawa  
The Univ. of Tokyo

1. Task

We want to know which words this learner knows.  
Application: reading support

Vocabulary Prediction

References:

Merging solves the problems a), b):  
a) merge graphs from multiple corpora  
b) merge graphs from corpora and graphs representing membership of words in a domain.

2. Current Method

1. Fix a corpus.  
2. Rank words in the corpus in descending order of its frequency.  
3. Tune ranking heuristically and manually (especially easiest words)

4. Group words by 1.000 words

5. Randomly sample 10 words from each level

Problems:  
a) Cannot handle multiple corpora directly  
b) Cannot create domain-specific test

3. Proposed Framework

For Prediction:  
Label Propagation [Zhou et al., 2004]

Repeatedly propagate labels of the nodes to their connected nodes

INPUT: a weighted graph, labeled nodes for training labeled nodes

OUTPUT: labels of the rest of the nodes (i.e., unlabeled nodes)

Cluster Assumption:  
A cluster of nodes connected heavily each other have similar labels.

labeled nodes

unlabeled nodes

Nodes: words  
In this task: Labeled: know/don’t know

How to determine seed nodes?

For Sampling:  
Non-interactive graph-based active learning [Ji et al., 2012] [Gu and Han, 2012]

INPUT: a weighted graph ONLY: seed nodes

OUTPUT: sampled words

Intuitive workflow of this algorithm:

1) Choose representative nodes in a cluster  
2) With avoiding sampling from neighbors of previously chosen nodes.

Numbers show the order in which nodes are sampled.

Default classifier of Gu and Han’s algorithm: LLGC (a label propagation Method by Zhou et al. 2004)

Merging solves the problems a), b):  
a) merge graphs from multiple corpora  
b) merge graphs from corpora and graphs representing membership of words in a domain.

4. Results

80%  
75%  
70%  
65%  
60%  
55%  
50%  
45%  
40%  
35%  
30%  
25%  
20%  
15%  
10%  
5%  
0%

Accuracy (%)  
Size of Training Samples

Classifiers: LLGC is used unless specified by ()

a) Enabled use of multiple corpora increased accuracy.

b) A test specific to computer domain was successfully created without decreasing accuracy over general words. Thus, we can measure both general and domain-specific vocabulary of learners.

References:

Quanquan Gu and Jiawei Han. 2012. Towards active learning on graphs: An error bound minimization approach. In Proceedings of ICDM 2012.
