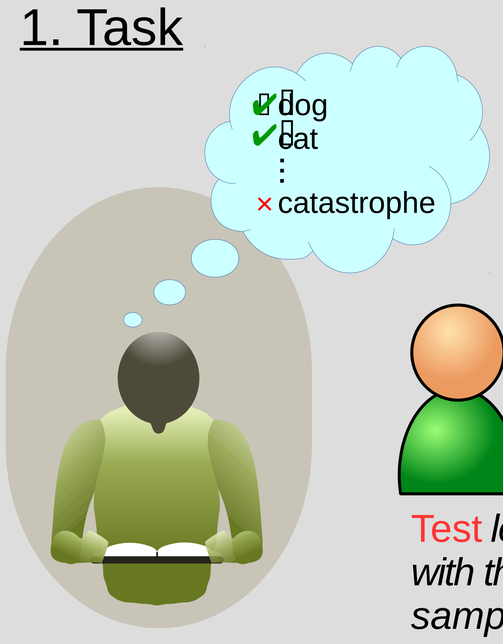
Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning

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 know.
Application: reading support
[Ehara et al., ACM TIST 2013]

Vocabulary Prediction [Ehara et al., COLING 2012]

Sample *words from the entire vocabulary*

Test *learners with the sampled words*

This work proposed *new framework* for here

Predict *the remainder of vocabulary using sampled words as training data*

2. Current Method

[Meara and Buxton, 1987], [Nation, 2007]

1. Fix a corpus.
2. Rank words in the corpus in descending order of its frequency.

the be of ahead cat… catastrophe …

3. Tune ranking heuristically and manually (especially easiest words)

the of cat be ahead … catastrophe …

4. Group words by 1,000 words

the of cat be ahead … catastrophe …

Level 1 Level 2 Level 3 Level 4 …

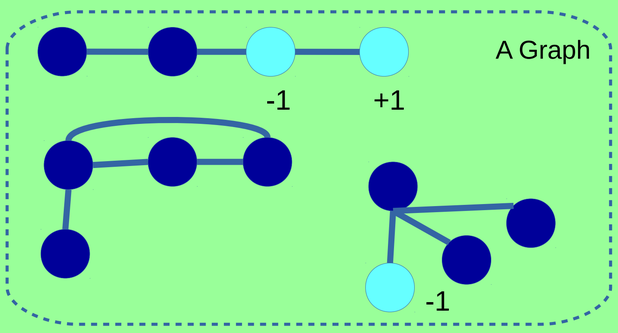
5. Randomly sample 10 words from each level

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)



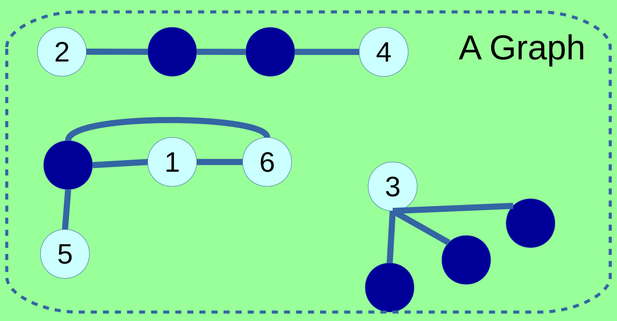
Cluster Assumption:
A cluster of nodes connected heavily each other have similar labels.

labeled nodes
unlabeled nodes

In this task:. Nodes: words
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*
OUTPUT: seed nodes

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)

Contribution:

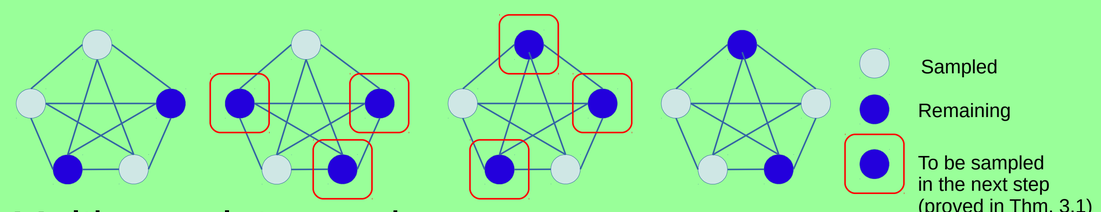
We formalized the current method as graph-based active learning problem. This formalization enables extending graphs so that problems a) and b) be solved.

Generalization

Making groups of *n* *words* and sample *k* *words* *from each*
=Making complete graphs of *n* *nodes* and sample *k* *nodes* *from each*

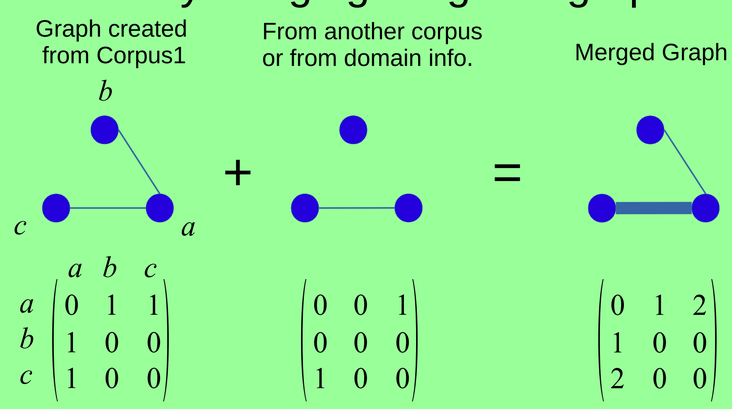
the of cat be ahead … catastrophe …

Level 1 Level 2 Level 3 Level 4 …



Multi-complete graph

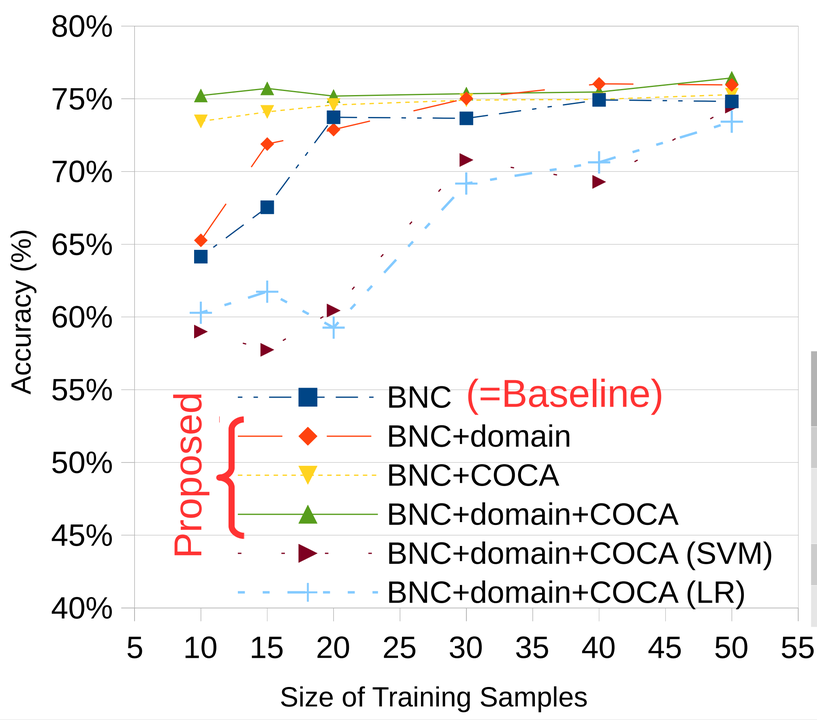
Extension: by merging weighted graphs sharing nodes



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.

BNC	Current method. Baseline. Multi-complete graph created from BNC corpus.
COCA	Multi-complete graph created from COCA corpus.
domain	We extracted all words in computer domain from the WordNet 3.0 and transformed the words into a complete graph.

4. Results



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.

Name	# of samples	Examples
BNC	0	-
BNC+domain	5	Input, client, field, background, register
BNC+COCA	0	-
BNC+domain+COCA	3	drive, client, command

Classifiers: LLGC is used unless specified by ().

References:
Quanquan Gu and Jiawei Han. 2012. Towards active learning on graphs: An error bound minimization approach. In *Proceedings of ICDM* 2012.
Dengyong Zhou, Oliver Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Scholkopf. 2004. Learning with local and global consistency. In *Proceedings NIPS*, pages 321– 328.
Batia Laufer and Paul Nation. 1999. A vocabulary- size test of controlled productive ability. *Language testing*, 16(1):33–51.