

Data Selection in Semi-supervised Learning for Name Tagging

Heng Ji and Ralph Grishman

(hengji, grishman)[@cs.nyu.edu](mailto:cs.nyu.edu)

July 2006

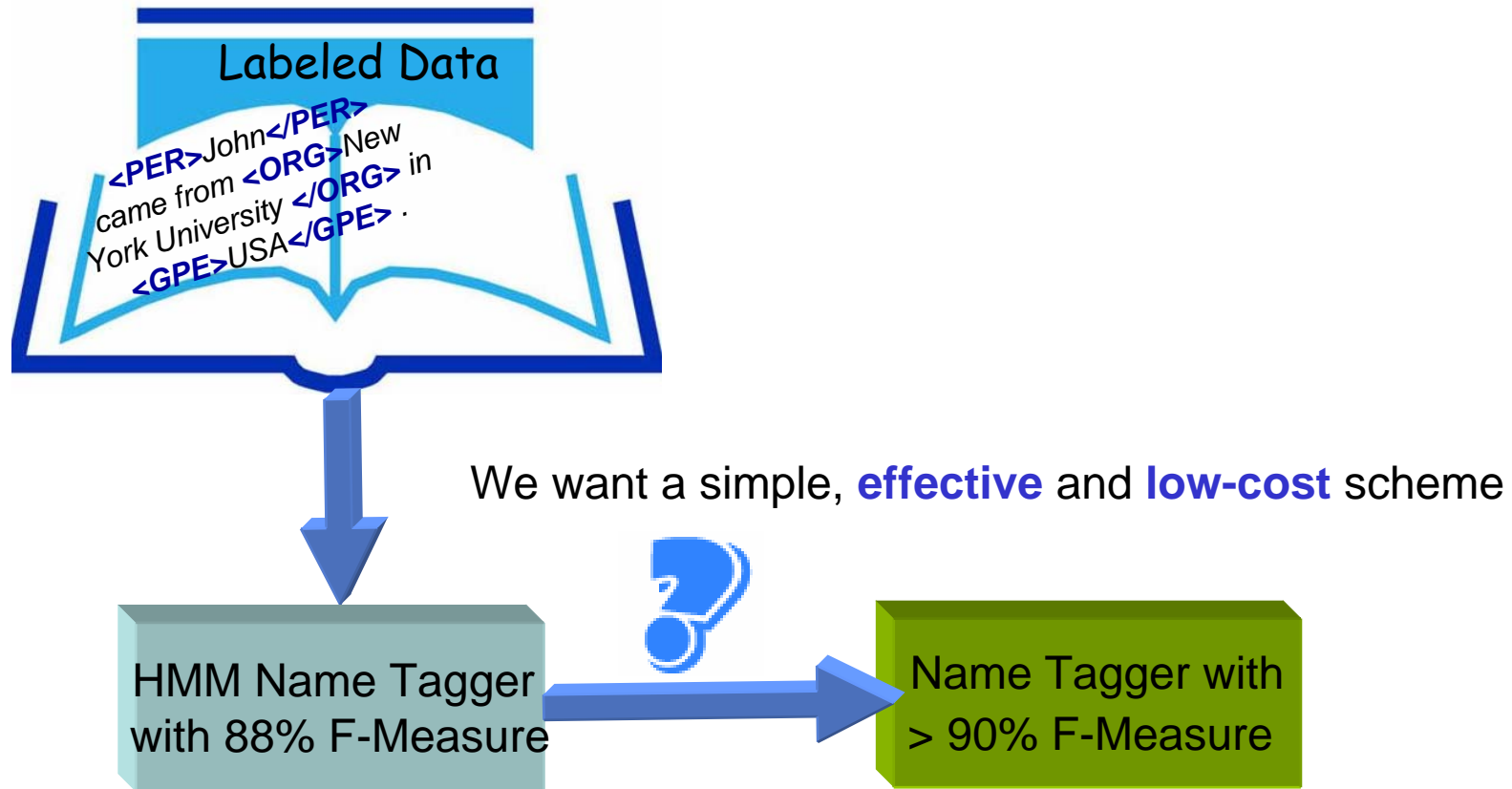


Outline

- Problem in Using Very Very Large Unlabeled Data
- Prior Work
- Two Semi-supervised Learning Methods for Name Tagging
 - Bootstrapping
 - Self-Training
- Unlabeled Data Selection
 - Document Selection
 - Sentence Selection
- Experimental Results
- Conclusions and Future work

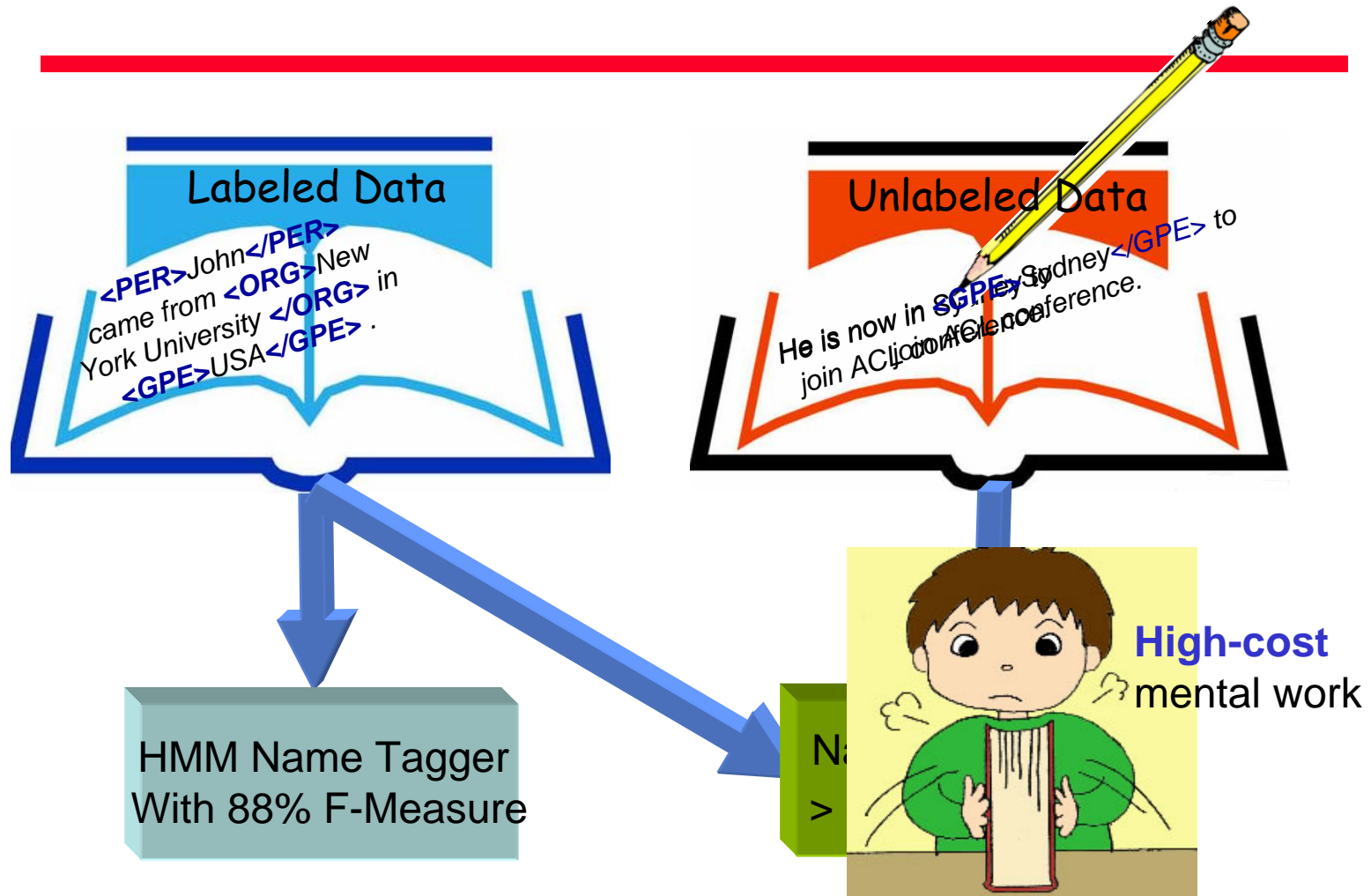


Goal: To Boost a Good-Performance Name Tagger



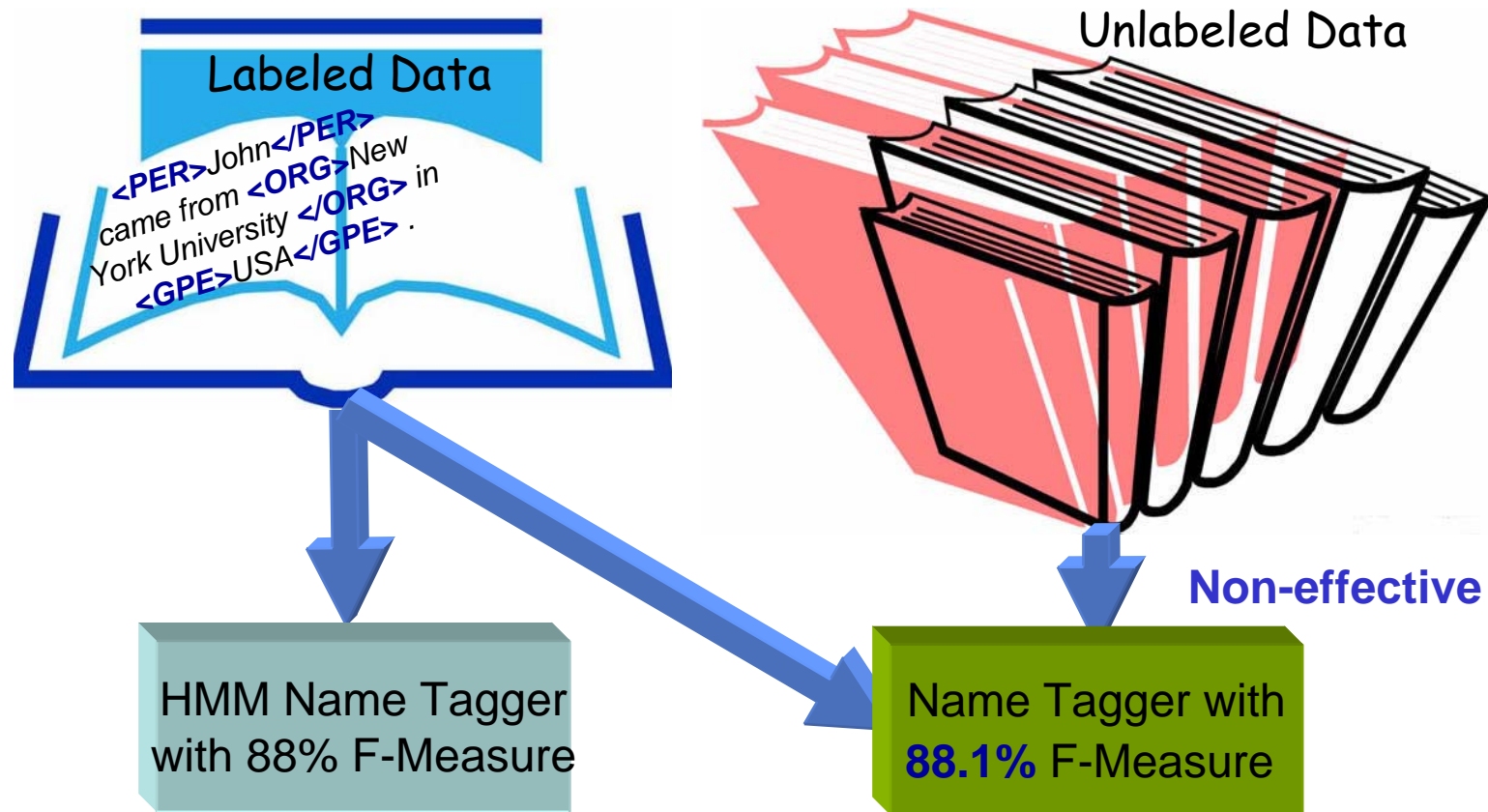


Solution 1: Hire Human Annotator ?





Solution 2: Add VERY VERY Large Unlabeled Data Directly?





Prior Work

- Very large data always helps
 - Banko and Brill (2001): a logarithmic trend in performance as corpus size increases without performance reaching an upper bound
 - Thesaurus extraction (Curran and Moens, 2002)
 - Is-a relation extraction (Ravichandran et al., 2004)

- Very large data does not always help
 - (Curran, 2002) and (Curran and Osborne, 2002): the choice of statistical model is more important
 - Our work: **Data selection is more important**



Our General Solution

- Where to get the unlabeled data
 - Extra large unlabeled data → Bootstrapping
 - Test set itself → Self-Training

- How to select uniformly 'useful' data
 - Clean and reliable
 - Relevant to test set
 - Automatically select



Semi-supervised Learning for Name Tagging

➤ Motivation

The text in which the tagger has low confidence may get ‘support’ from those texts previously labeled with high confidence

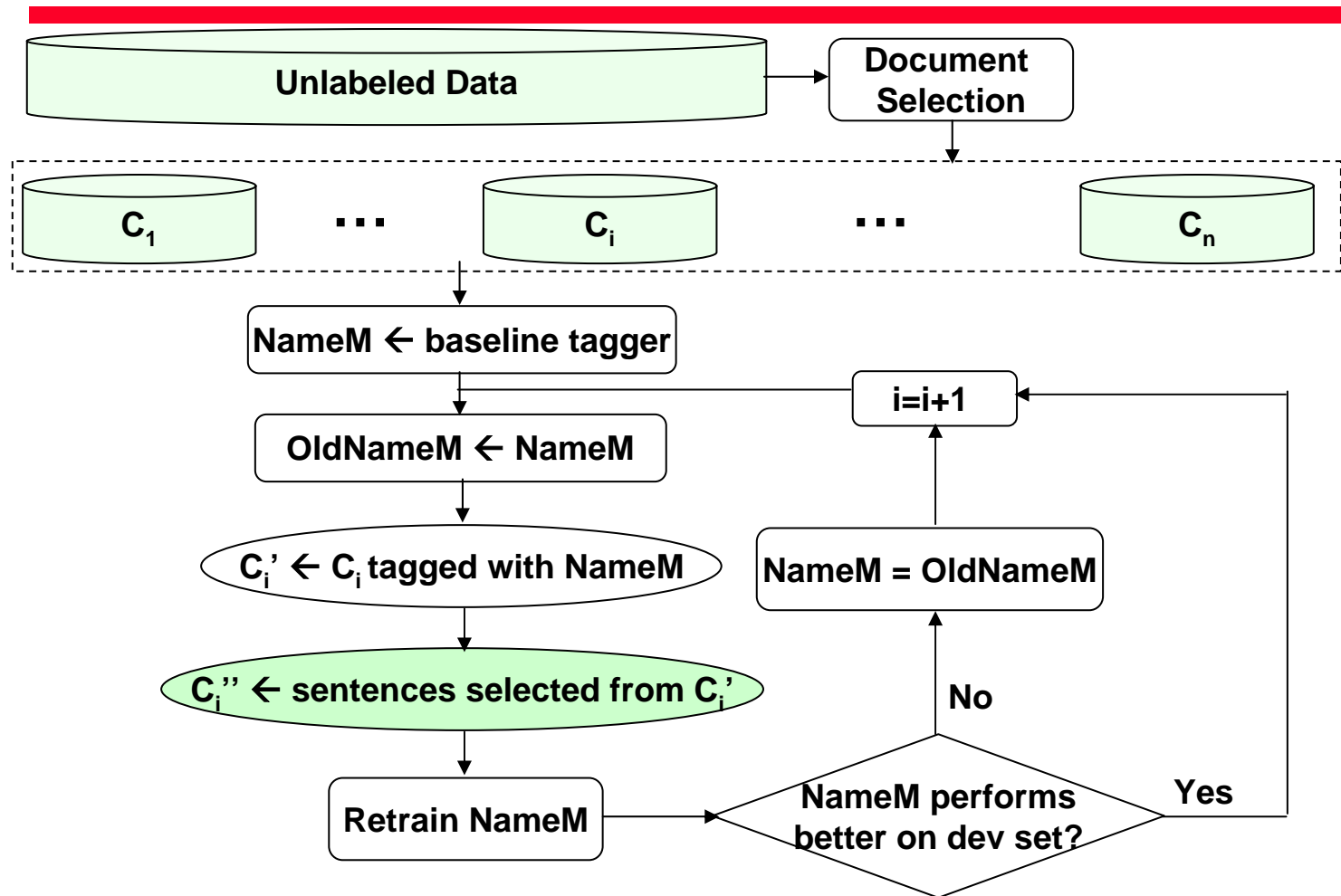
- Learn a previously-unseen token tagged with high confidence as a predictive token
- Learn a previously-unseen context that appears consistently as a predictive context

➤ Iterative Procedure

- Take the unlabeled text about which the tagger has greatest confidence in its decisions, tag it
- Add the tagged text to the training set
- Re-train the tagger

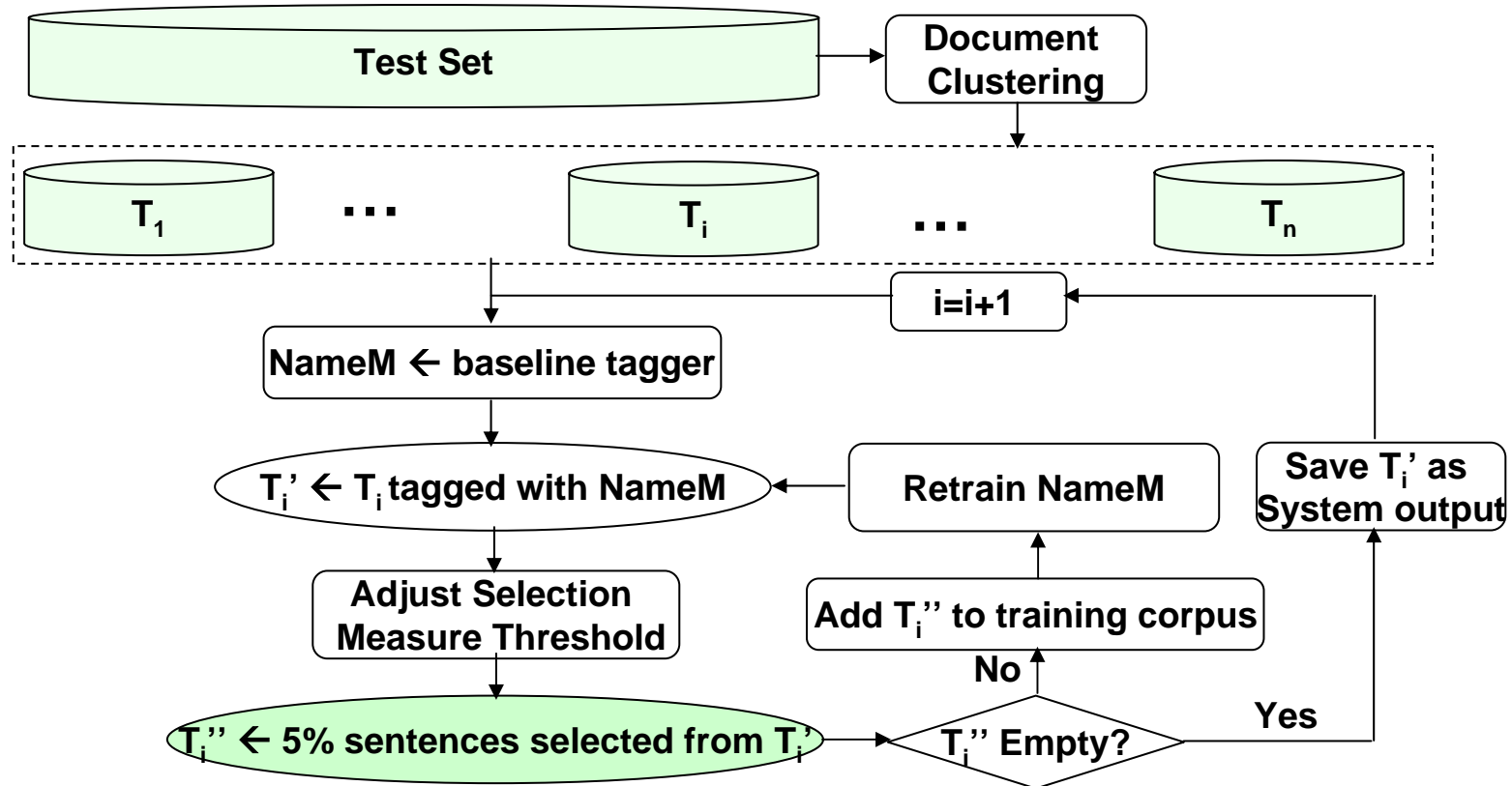


Bootstrapping for Name Tagging





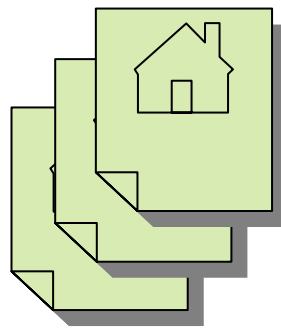
Self-Training for Name Tagging





Select Relevant Documents

Test Set



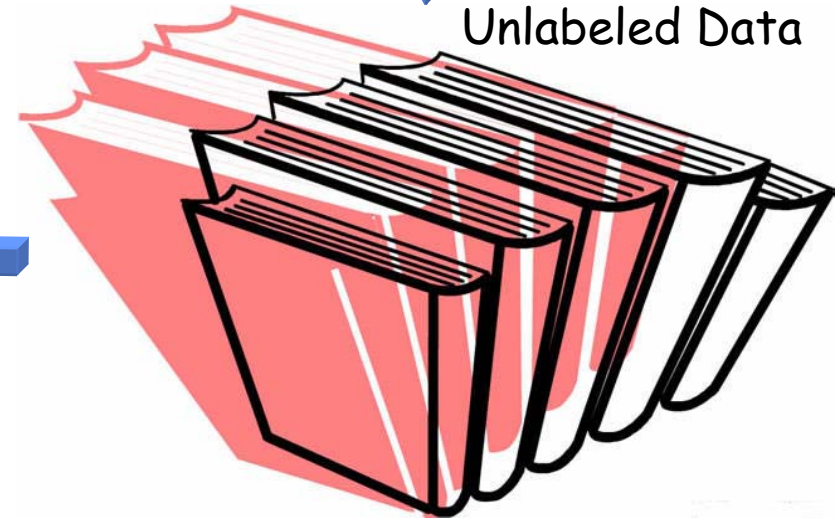
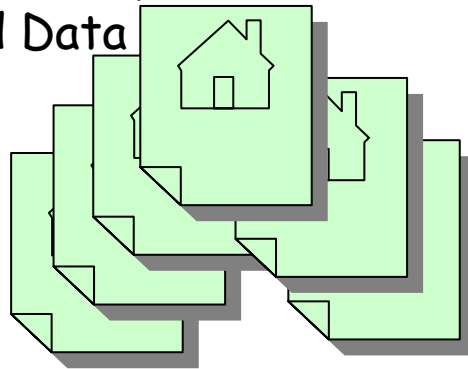
Baseline
Name Tagger

Query Set Q

Name Candidates
from N-Best lists

Unlabeled Data

Useful ('Relevant')
Unlabeled Data





Cross-entropy to Count Document Relevancy

➤ Definition

$$H(\text{TestSet}, d) = - \sum_{x \in Q} \text{prob}(x | \text{TestSet}) \times \log_2 \text{prob}(x | d)$$

- Q : the name candidates in the top N best lists of the sentences in the test set
- x : a name candidate in the Q
- d : a document in the unlabeled corpus

➤ Usage

- In bootstrapping: to select documents in the unlabeled corpus
- In self-training: to cluster documents in the test set



Select Reliable and Informative Sentences

➤ Margin

- Definition: the difference in log probabilities between the first hypothesis and the second hypothesis
- The sentences with high margins are a mix of interesting and uninteresting samples →

➤ Name Coreference

- k names tagged in sentence S : $\{N_1, N_2, \dots, N_k\}$, which are coreferred to by $\{\text{CorefNum}_1, \text{CorefNum}_2, \dots, \text{CorefNum}_k\}$ other names, count:

$$\text{AveCoref}_S = \left(\sum_{i=1}^k \text{CorefNum}_i \right) / k$$

- Select S with large AveCoref_S

➤ Name count and sentence length

- Don't use a sentence if it includes fewer than six words, or doesn't include any names



Experiments: Baseline System

- A multiple-hypotheses HMM name tagger to process English and Chinese documents
- Name types: Person (PER), Organization (ORG) and Geo-political (GPE)



Experiments: Data

Data		English	Chinese
Baseline Training data		ACE02,03,04 989,003 words	Beijing Corpus +ACE03,04,05 1,460,648 words
Unlabeled Data	Total	196,494 docs in Mar-Jun of 2003 (69M words) from ACE05 unlabeled data	41061 docs in Nov,Dec of 2000, and Jan of 2001 (25M words) from ACE05 and TDT4 transcripts
	Selected Docs	62584 docs (1,314,148 Sentences)	14,537 docs (222,359 sentences)
	Selected Sentences	290,973 sentences (6,049,378 words)	55,385 sentences (1,128,505 words)
Dev Set		20 ACE04 texts in Oct of 2000	90 ACE05 texts in Oct of 2000
Test Set		20 ACE04 texts in Oct of 2000 and 80 ACE05 texts in Mar-May of 2003 (3093 names, 1205 PERs, 1021GPEs, 867 ORGs)	90 ACE05 texts in Oct of 2000 (3093 names, 1013 PERs, 695 GPEs, 769 ORGs)



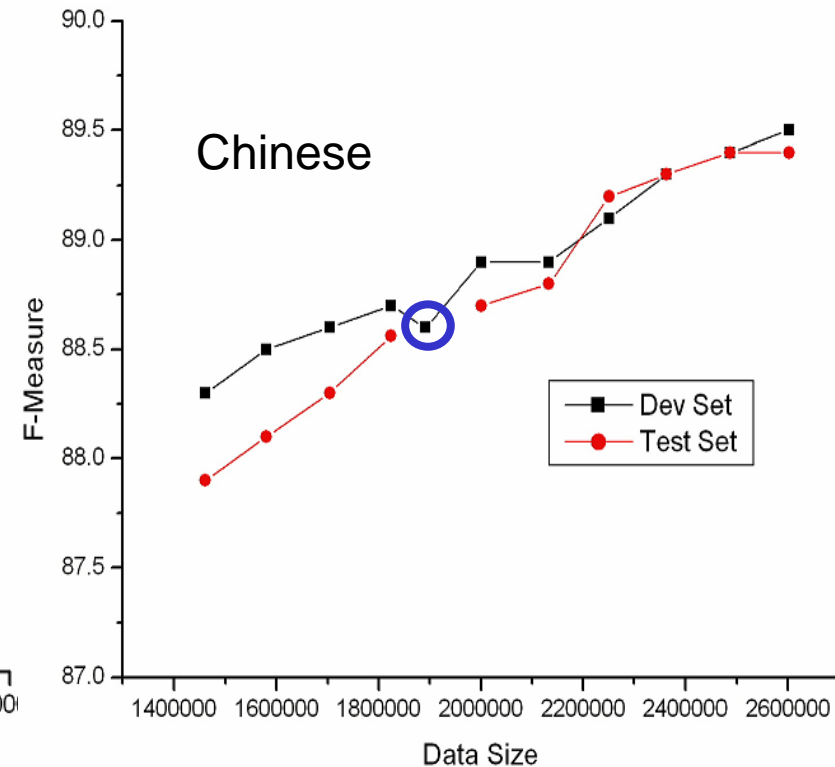
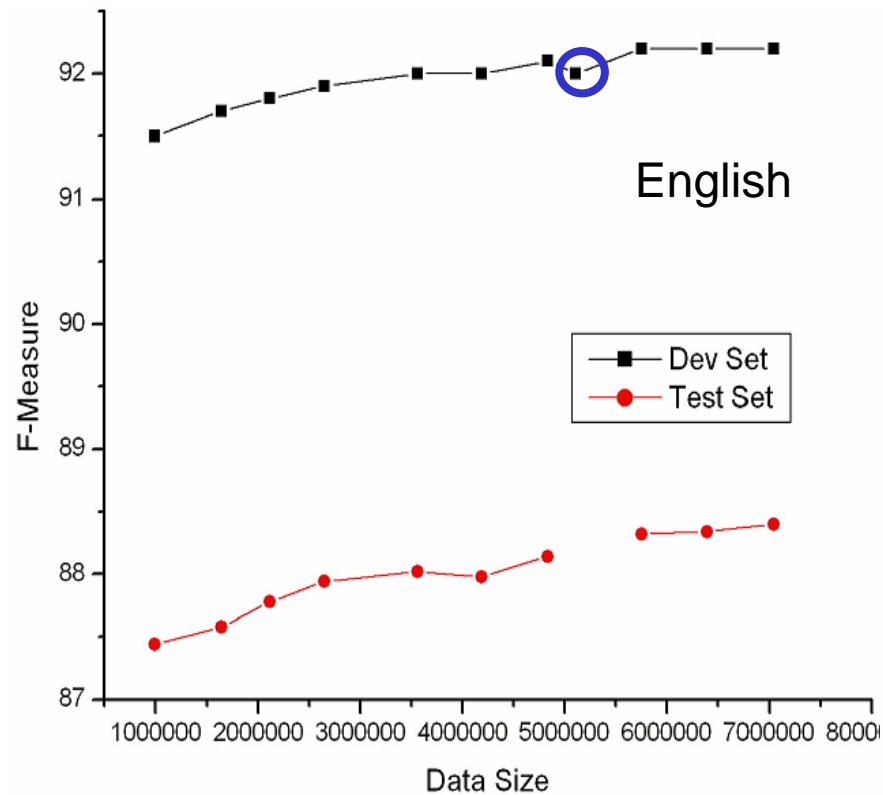
Experiment: Overall Performance

Learner	English			Chinese		
	P	R	F	P	R	F
Baseline	87.3	87.6	87.4	88.2	87.6	87.9
Bootstrapping with data selection	88.2	88.6	88.4	89.8	89.5	89.6
Self-training	88.1	88.4	88.2	89.5	88.3	88.9
Bootstrapping with data selection + Self-training	89.0	89.2	89.1	90.2	89.7	90.0

➤ Passed significance testing

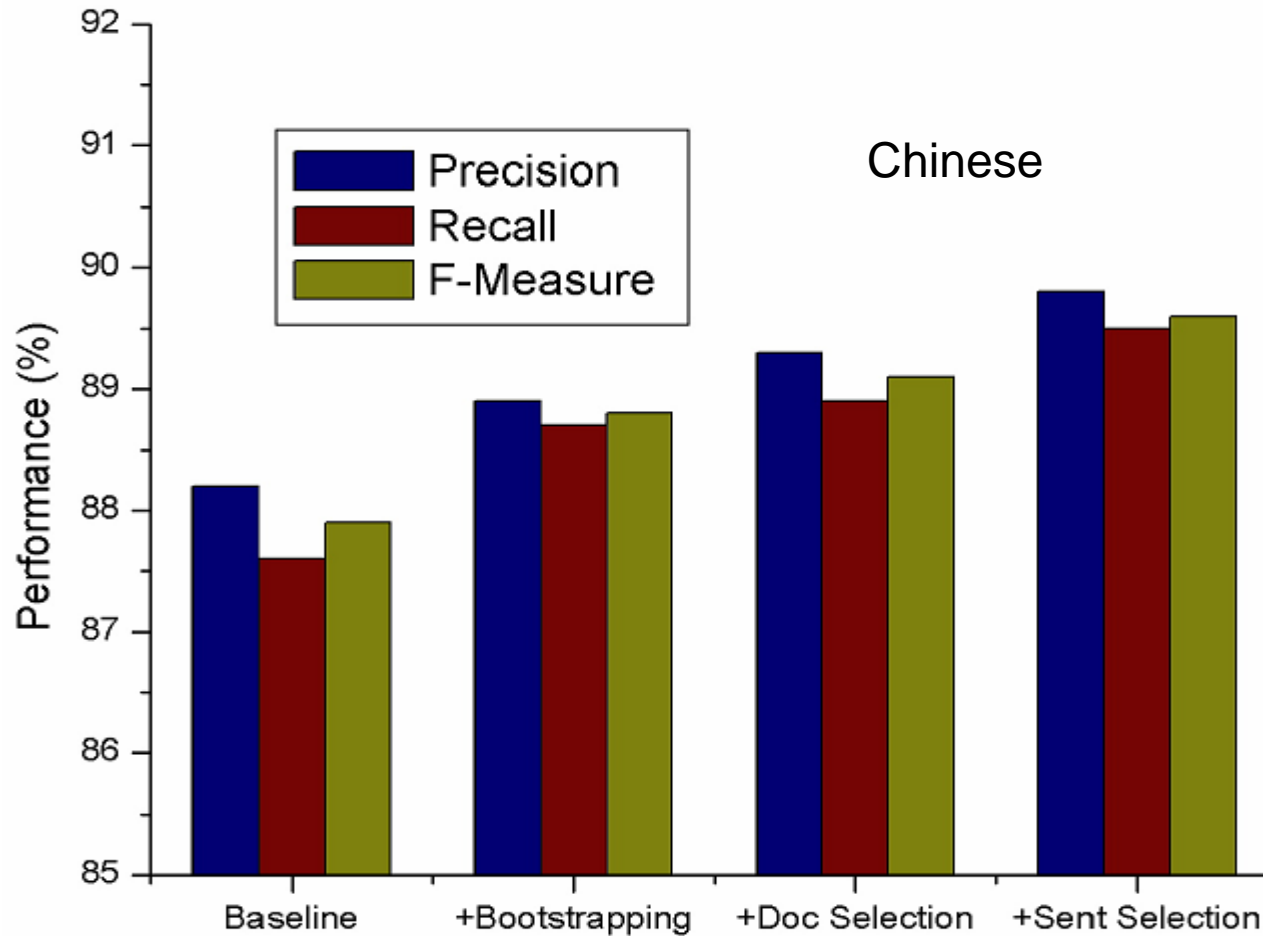


Experiments: Impact of Data Size in Bootstrapping



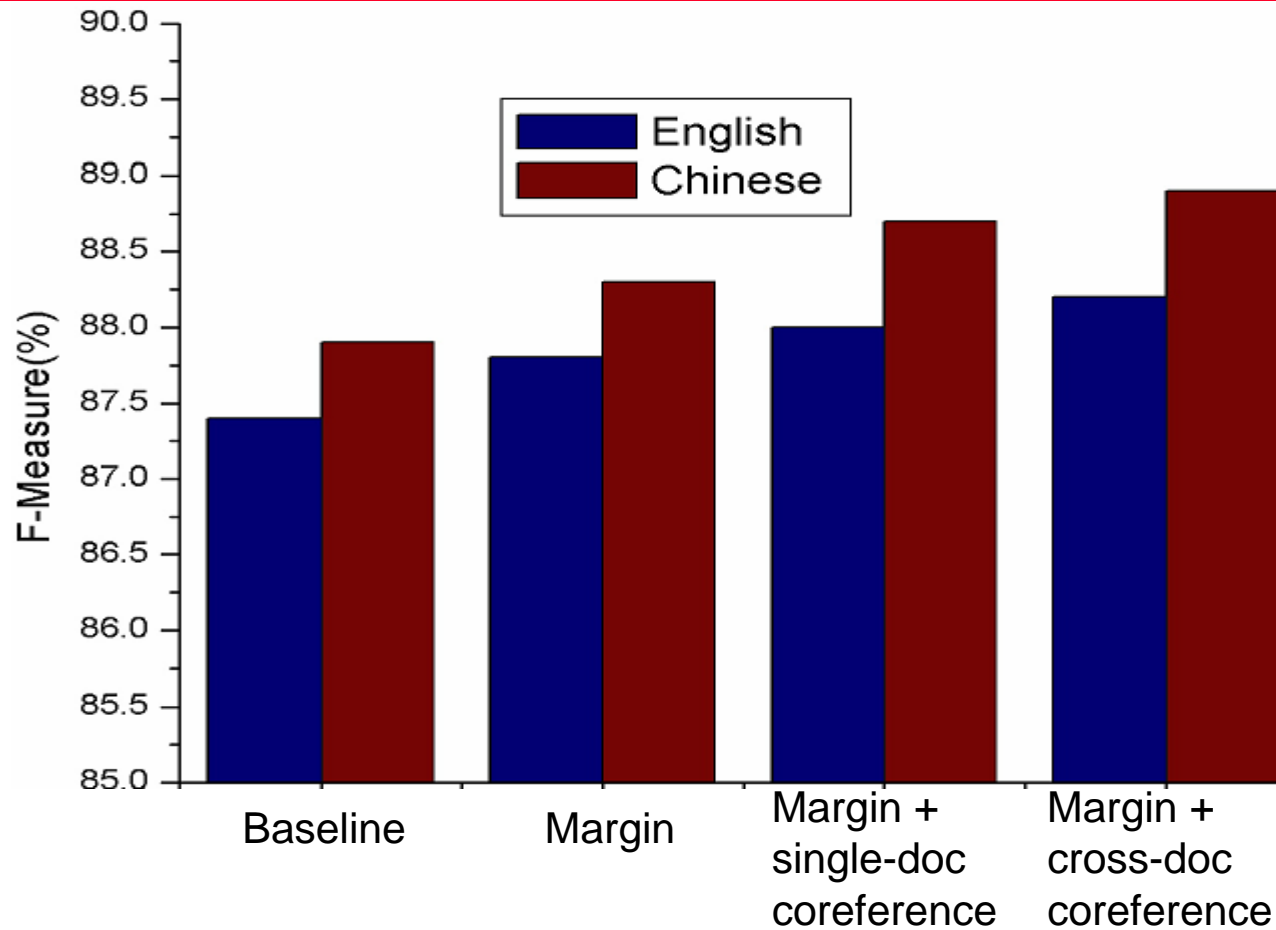


Experiments: Impact of Data Selection in Bootstrapping





Experiments: Impact of Sentence Selection in Self-Training





Experiments: Adjust Margin Threshold in Self-Training

No. of Iterations	Number of Sentences added	Number of Tags added	F-Measure
0	0	0	91.4
1	37	28	91.9
2	69	22	92.1
3	107	21	92.4
4	128	11	92.6
5	146	9	92.7
6	163	8	92.8
7	178	6	92.8
8	190	0	92.8

- Test on one English cluster including 7 documents and 190 sentences
- Lower the margin threshold so that about 5% of the sentences (with the largest margin) are added to the training corpus



Conclusions and Future Work

- Demonstrated the effectiveness of two semi-supervised learning methods for name tagging
- Investigated the importance of data selection
Exploiting a very large corpus did not by itself produce the greatest performance gain
- We select “easy” sentences to add automatically to the training set, which can be combined with active learning approach (select “difficult” sentences + label manually)
- Future Work
 - Try topic identification techniques to select relevant unlabeled documents
 - Use the downstream IE components to measure tagging confidence
 - Apply clustering as a pre-processing step for bootstrapping



Thanks!

