Lexicon Acquisition for Dialectal Arabic using Transductive Learning

Kevin Duh and Katrin Kirchhoff

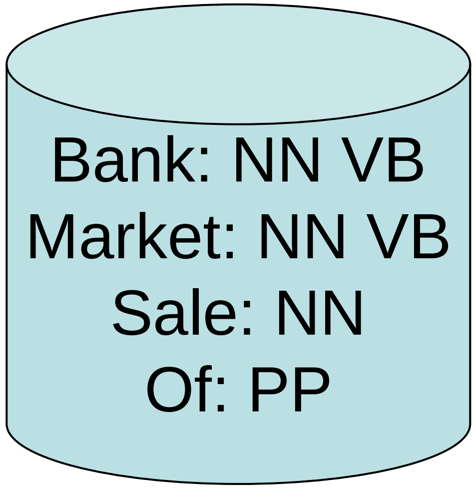
University of Washington

Motivation

- Motivation:
  - Develop NLP tools/applications for resource-poor languages
- Resource-poor languages
  - Lack annotated data (lexicon, treebank, labeled text)
  - Examples: Arabic dialects, languages of India, China
- Current supervised NLP methods are not adequate for resource-poor languages
  - Too much reliance on availability of annotated data

This work

- Learning a POS lexicon for dialectal Arabic (a resource-poor language)



- Why POS lexicon?
  - Essential resource in unsupervised tagging
  - POS tagging is first step to many NLP systems

Contributions

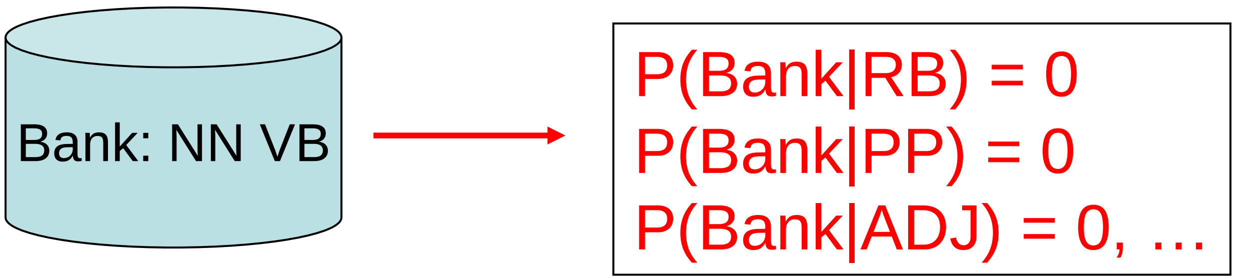
1. Problem formulation: Lexicon acquisition as transductive learning
2. Comparison of 3 transductive learning algorithms
   - Transductive SVM
   - Spectral Graph Transducer
   - Transductive Clustering
3. Demonstrate tagging improvement in dialectal Arabic

Why is the lexicon important in unsupervised tagging?

- HMM tagger

$$p(word_{1:N}, tag_{1:N}) = \prod_{i=1}^{N} p(word_i \mid tag_i)\, p(tag_i \mid tag_{i-1})$$

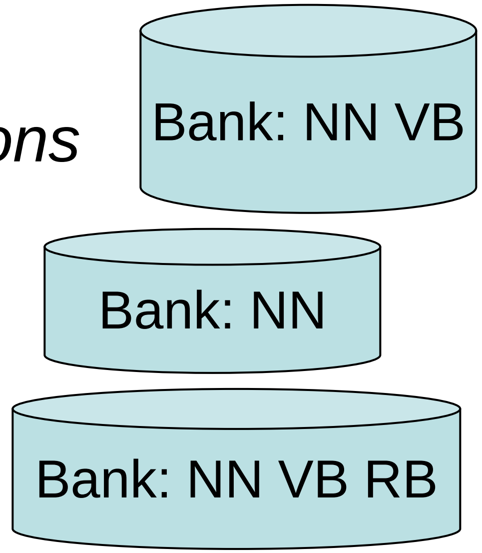
  - EM: Adjust parameters to maximize likelihood on raw text (many local optima)
- Lexicon adds knowledge to $p(word_i|tag_i)$, $p(tag_i|tag_{i-1})$
  - E.g.



  - These zero probabilities add hard constraints and biases EM to avoid certain solutions

Difference between good and bad lexicons is drastic

- A **good** lexicon:
  - *Reduces parameter space,*
  - *Guides EM to better predictive distributions*
- A **poor** lexicon:
  - *May never hypothesize correct tag*
  - *May result in bad local optimum for EM*



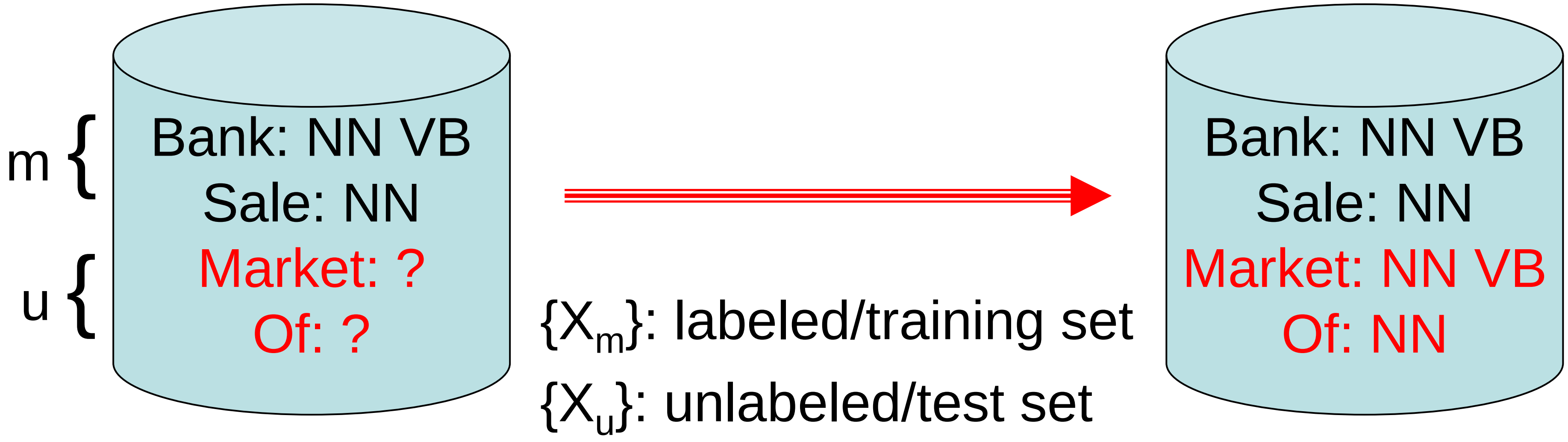
- English WSJ Results[Banko&Moore'04][Wang&Schuurmans'05]
  - If lexicon doesn't filter low frequency tags, unsupervised tagger accuracy decreases from 96% to 77%

Outline

1. Motivation & Importance of Lexicon in Unsupervised Tagging
2. Lexicon Learning
   a) Problem Formulation
   b) 3 Transductive Learning Algorithms
3. Experiments in Dialectal Arabic
4. Conclusions

Lexicon Learning: Problem Formulation

- How does one build a lexicon?
  1. Ask an expert to label all words, or collect labels from POS-annotated text (Resource-intensive!)
  2. Ask an expert to label some words, use machine learning to learn the rest (Scalable to amount of effort)

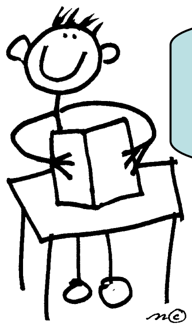


Task: Given $\{X_m\}$, predict labels of $\{X_u\}$ with low error

Lexicon learning is a transductive learning problem

	Transductive Learning	Inductive Learning
Goal	Label the test set, given during learning	Learn a function to label any future test set
Resource	1. Labeled training set 2. Unlabeled test set	Training set: (labeled,unlabeled,both) (supervised,un-/semi-supervised)
Suitable Problems	Test set is available & fixed	Test set is revealed in the future

Transductive learning = take-home exam



Inductive learning = in-class exam

m { Bank: NN VB
Sale: NN

u { Market: ?
Of: ?

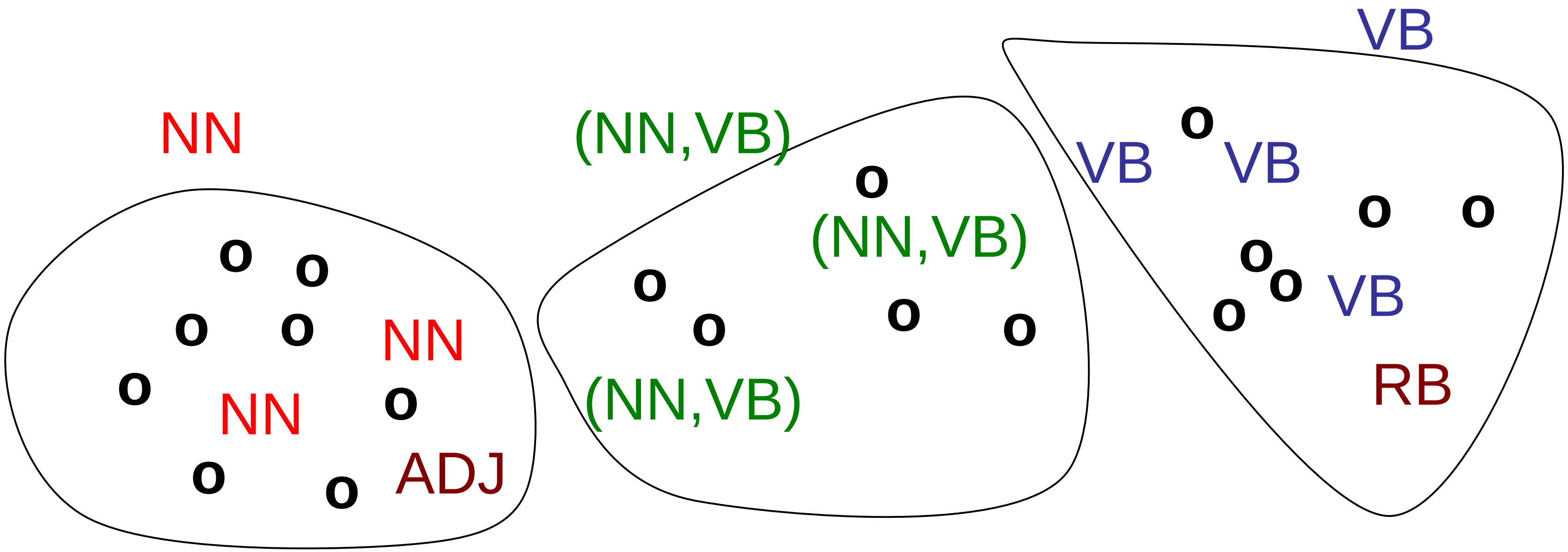
Next up:
3 Transductive Learning Algorithms

1. Transductive Clustering
2. Transductive SVM
3. Spectral Graph Transducers

A simple transductive algorithm

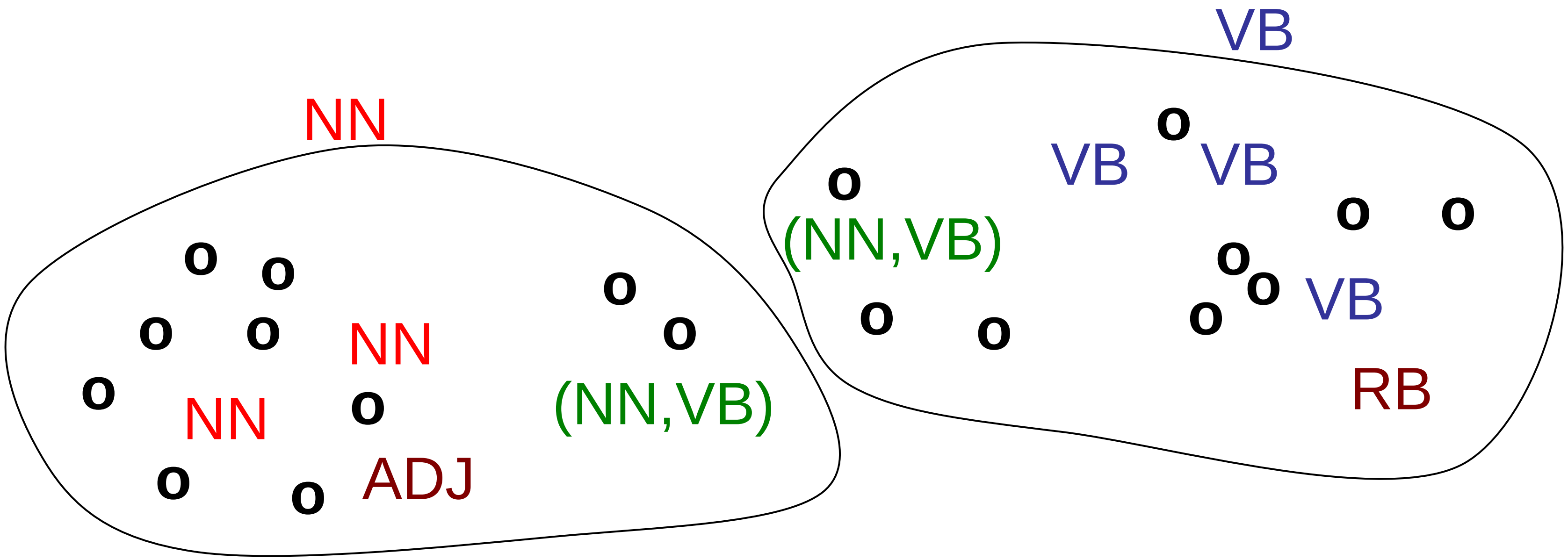
- Assumption: Samples close together have the same label
  - Corollary: Only 1 label is needed for all samples that form a cluster

- Basic algorithm:
  1. Cluster all data
  2. Label test samples with majority (plurality) label of cluster



A simple transductive algorithm

- Issue: How to decide the number of clusters?



Error bound

- Solution: Use an error bound to choose # of clusters (different hypotheses)
- [Derbeko et. al., JAIR'04] proved a bound for transductive learning:
  - With probability $1-\delta$, a hypothesis h has bound:

$$R_h(X_u) \le \hat{R}_h(X_m) + \sqrt{\left(\frac{m+u}{u}\right)\left(\frac{u+1}{u}\right)\left(\frac{\ln(1/p(h)) + \ln(1/\delta)}{2m}\right)}$$

Test Risk

Empirical Risk

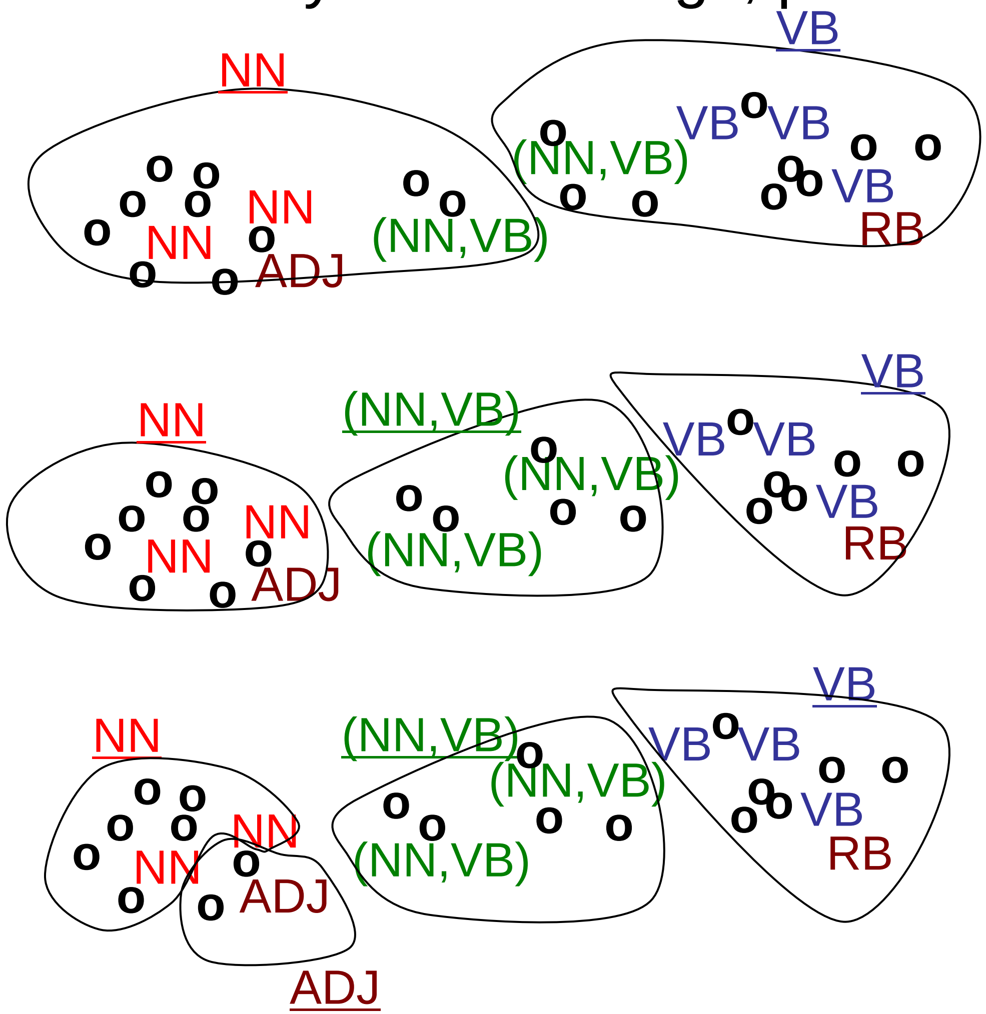
m: # labeled samples
u: # unlabeled samples

Prior probability of hypothesis h

A good hypothesis has low Empirical Risk and high Prior

Transductive Clustering [El-Yaniv, 2005]

Idea: Try all clusterings; pick the one with lowest bound



Hypothesis: 2 clusters

$$R_{h2}(X_u) \leq 0.43$$

Hypothesis: 3 clusters

$$R_{h3}(X_u) \leq 0.25$$

Hypothesis: 4 clusters

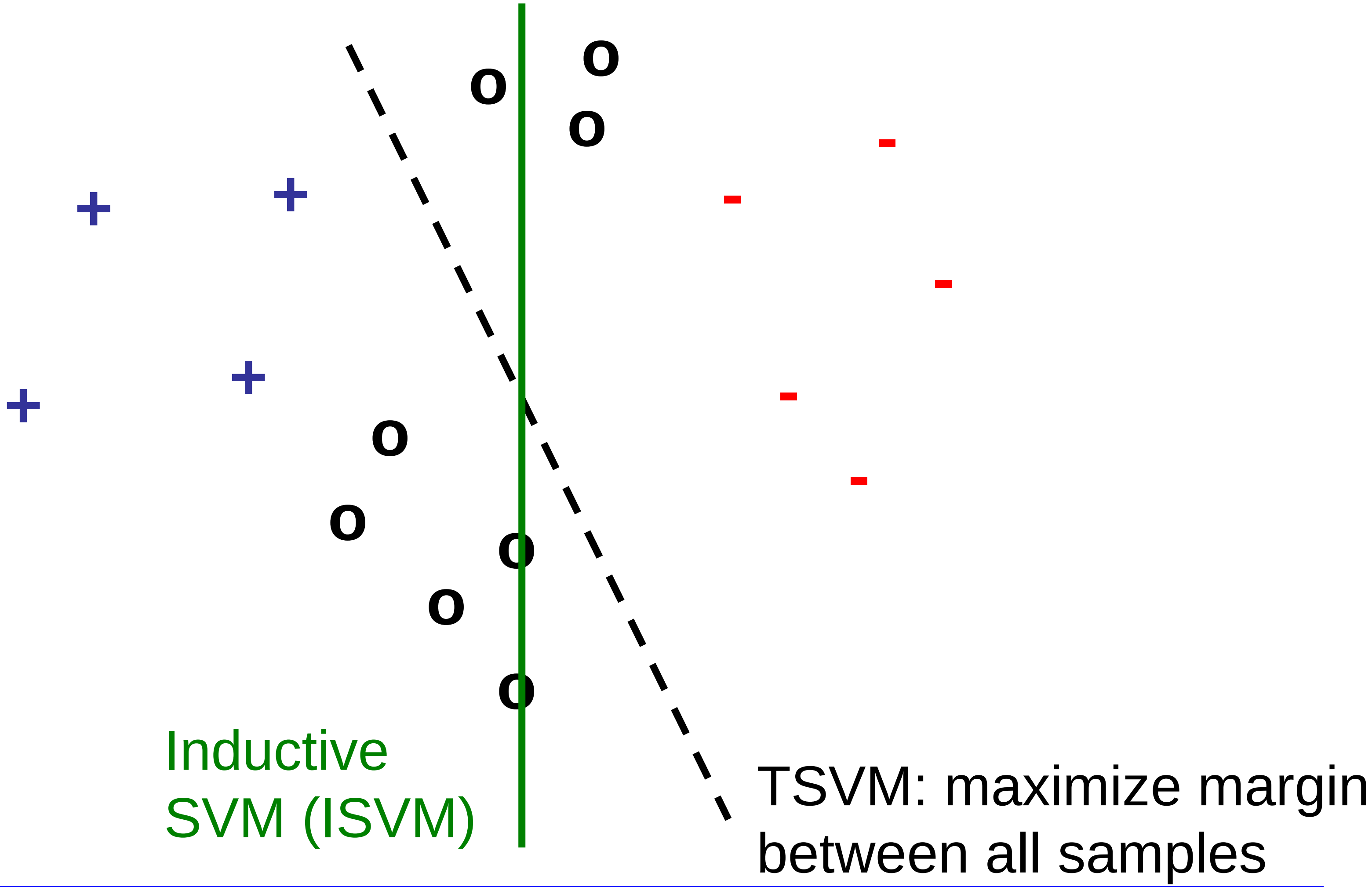
$$R_{h4}(X_u) \leq 0.32$$

Transductive Clustering: Pros & Cons

- Pros:
  - Theoretical guarantees
  - Easy to implement
  - Modular:
    - Use different clustering algorithms as input
  - No hyper-parameters - no tuning required
- Cons:
  - Accuracy is very dependent on cluster quality
    - But clustering may not be optimized for discrimination
  - Bound may be loose in large multi-class problems
    - A loose bound does not correlate well with test risk

$$R_h(X_u) \leq \hat{R}_h(X_m) + \sqrt{\left(\frac{m+u}{u}\right)\left(\frac{u+1}{u}\right)\left(\frac{\ln(1/p(h)) + \ln(1/\delta)}{2m}\right)}$$

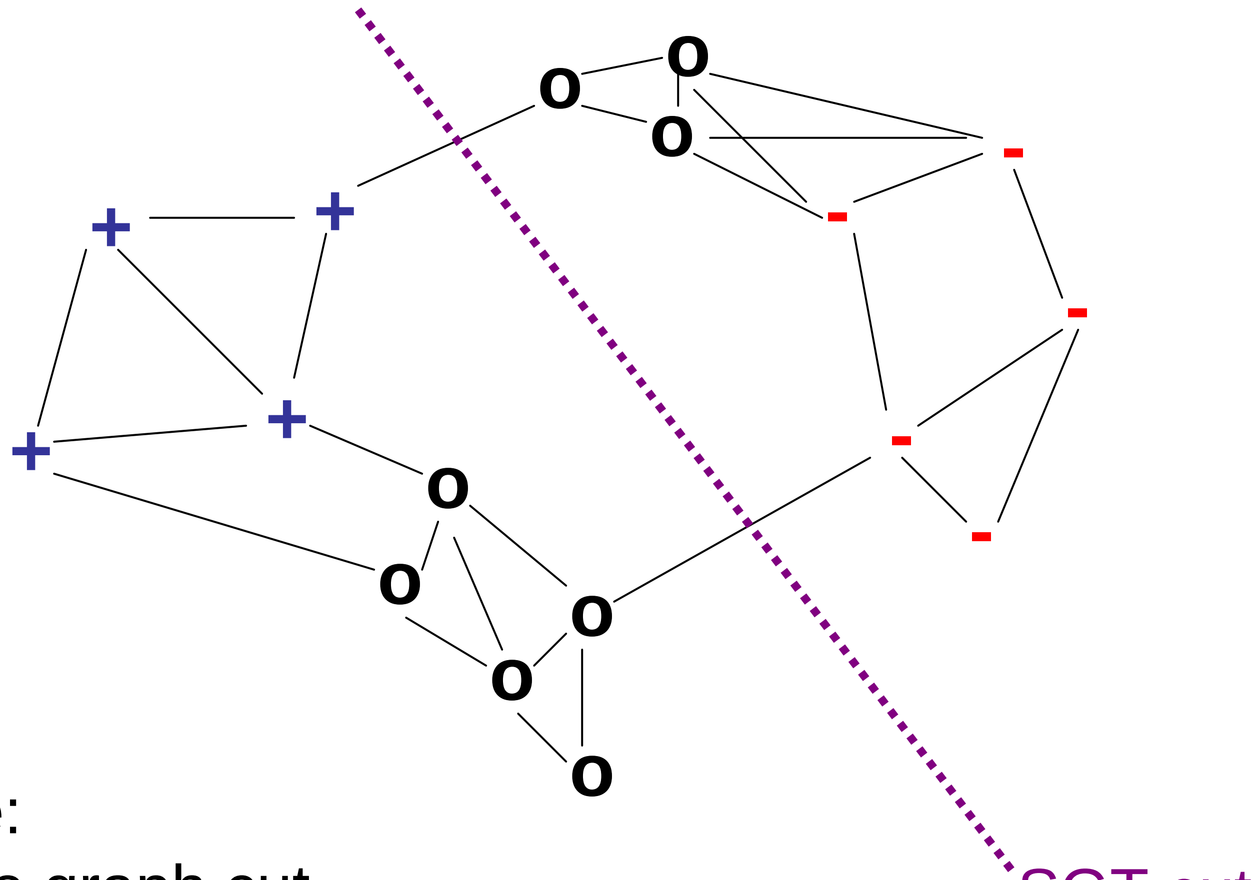
Transductive Support Vector Machines (TSVM) [Joachims, 1999]



Spectral Graph Transducer (SGT)

[Joachims, 2003]

Begin with a data graph that encode similarities between samples



Objective:

Minimize graph cut

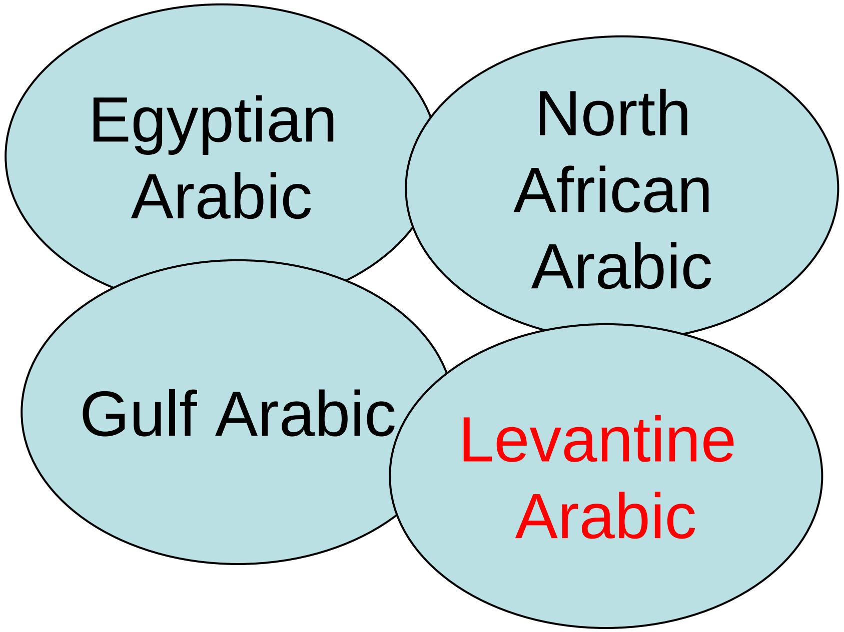
subject to constraints that labeled sample be in same cluster

Outline

1. Motivation & Importance of Lexicon in Unsupervised Tagging
2. Lexicon Learning
   a) Problem Formulation
   b) 3 Transductive Learning Algorithms
3. Experiments in Dialectal Arabic
   1. Available Resources
   2. Experimental Setup
   3. Results
4. Conclusions

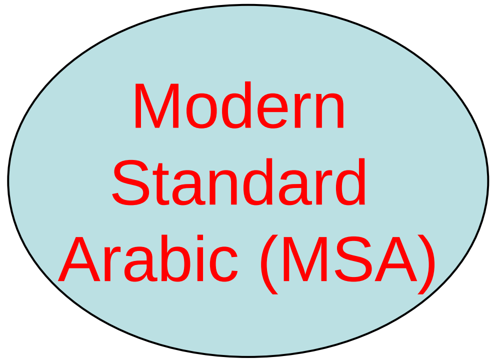
Dialectal Arabic and Available Resources

Spoken dialects: Everyday use



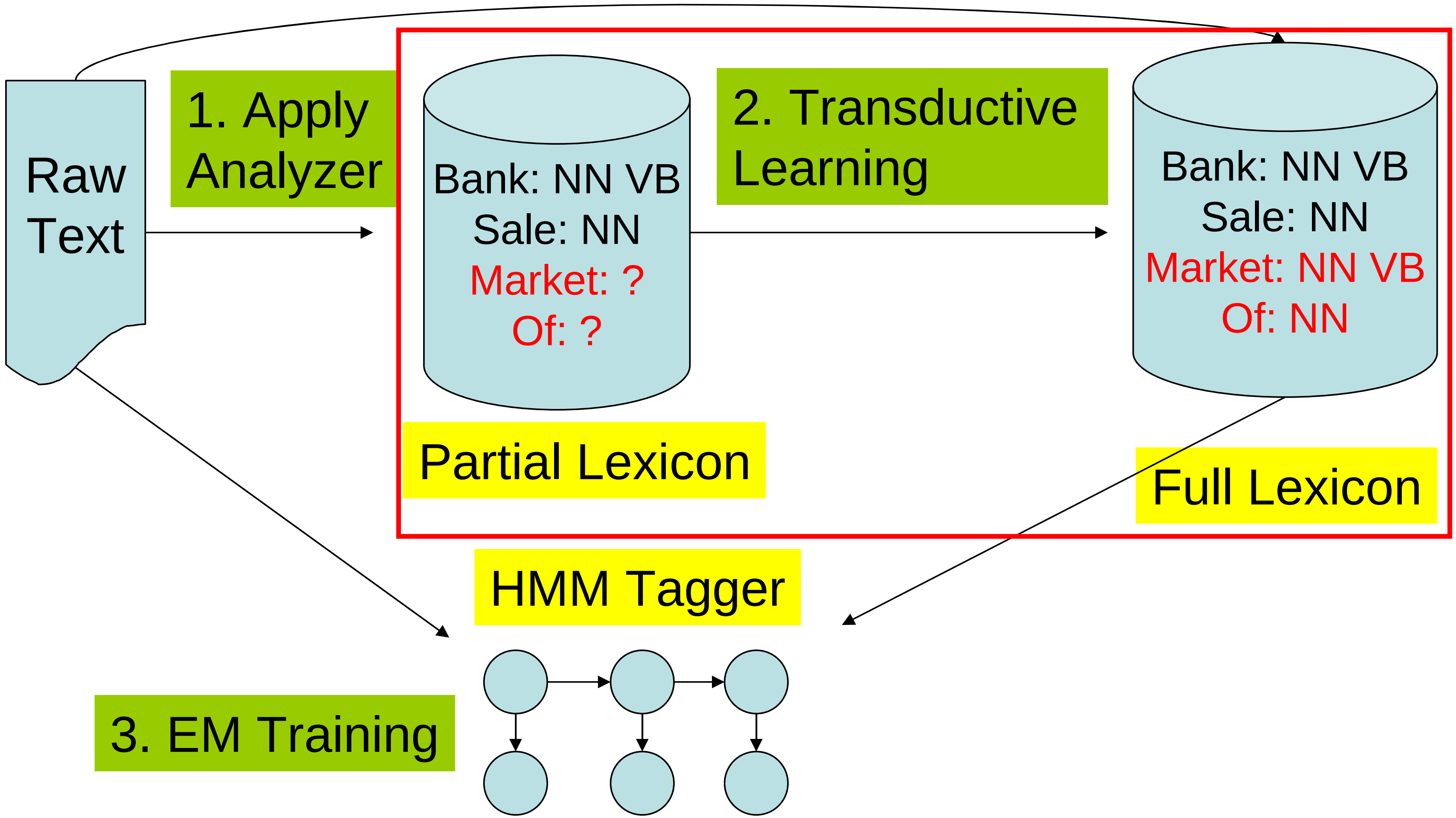
Levantine raw text (LDC CallHome)
- train unsupervised tagger
- wordlist for lexicon

Written, formal use

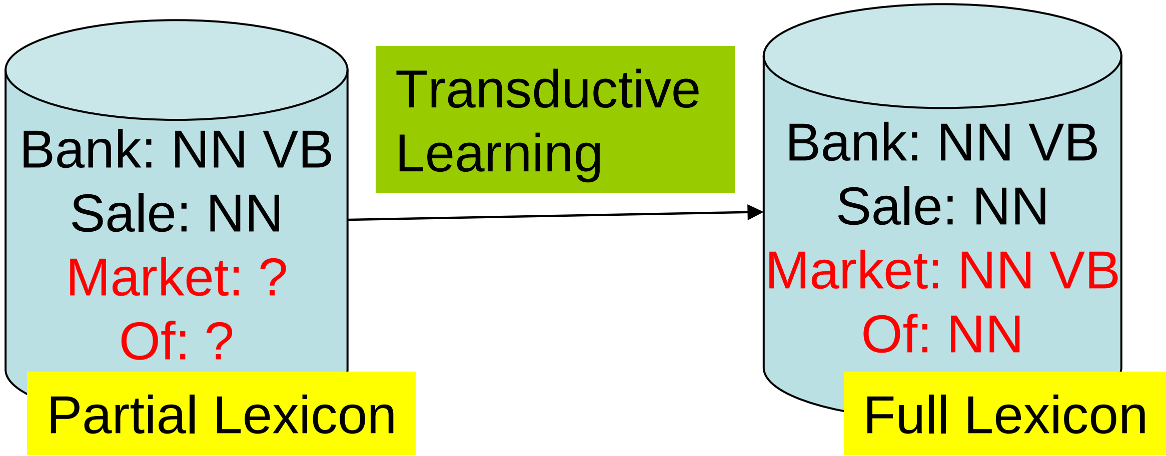


MSA Morphological Analyzer (by Buckwalter, LDC)
- labels some Levantine words

Experimental Setup



(Step 2) Lexicon Learning: Data & Features



- <u>Data:</u>
  - 23% of lexicon are unlabeled (4k of 16k words)
  - 20 tags in tagset, but ~200 labels (compound “NN-VB”)
- <u>Features (~17k features for each word):</u>
  - Orthographic: matching prefix/suffix
  - Contextual (counts from raw text):
    - Word bigram, POS bigram (if available)
  - All algorithms use same feature set

Results using taggers trained with different lexicons

Method for acquiring lexicon	Tag Accuracy
Baseline (All Tags)	55.6%
Baseline (Open Class)	57.4%
Spectral Graph Transducer	59.7%
Inductive SVM	61.5%
Transductive Clustering	**62.9%**
Transductive SVM	**63.5%**

Test set:
15k tokens
POS-annotated
(Levantine Arabic
CTS Treebank)

1. All machine-learned lexicons outperform baseline
2. Transductive Clustering & TSVM perform best:
   - both are transductive and have few hyperparameters

Conclusions

1. Lexicon acquisition as transductive learning
2. Compared 3 transductive algorithms
   - TSVM, SGT, Transductive Clustering
3. Results on Dialectal Arabic:
   - Using a machine-learned lexicon improves tagger accuracy (6% over baseline)
   - TSVM and Tranductive Clustering perform best

- <u>Future Work:</u>
  - Dealing with noisy expert labels
  - Improved Transductive Clustering
    - Semi-supervised clustering using labeled data
    - Error Bound for F-measure and other metrics

Thanks!

- Questions?

Comparison of Lexicons

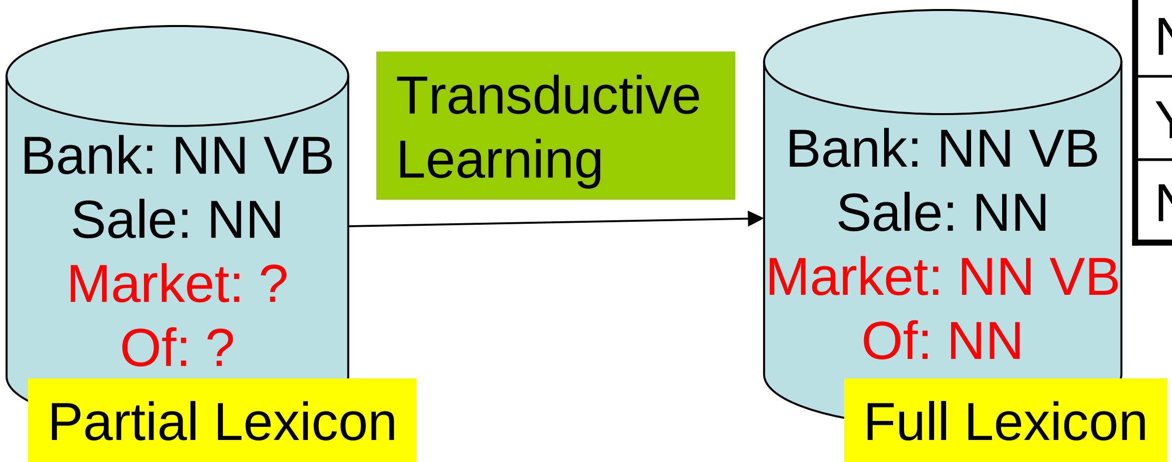
- 15k words in lexicon occur in Test Set
  - Collect “oracle” POS for these words as reference
  - Compute precision/recall of learned-lexicon

Method	Precision	Recall	POS size
TSVM	58.1	88.8	1.89
TC	59.2	87.9	1.80
ISVM	58.1	88.4	1.87
SGT	54.0	82.6	1.87
Open class	54.0	96.7	3.39
All tags	53.3	98.5	5.17

Error Propagation: Preliminary Evaluation

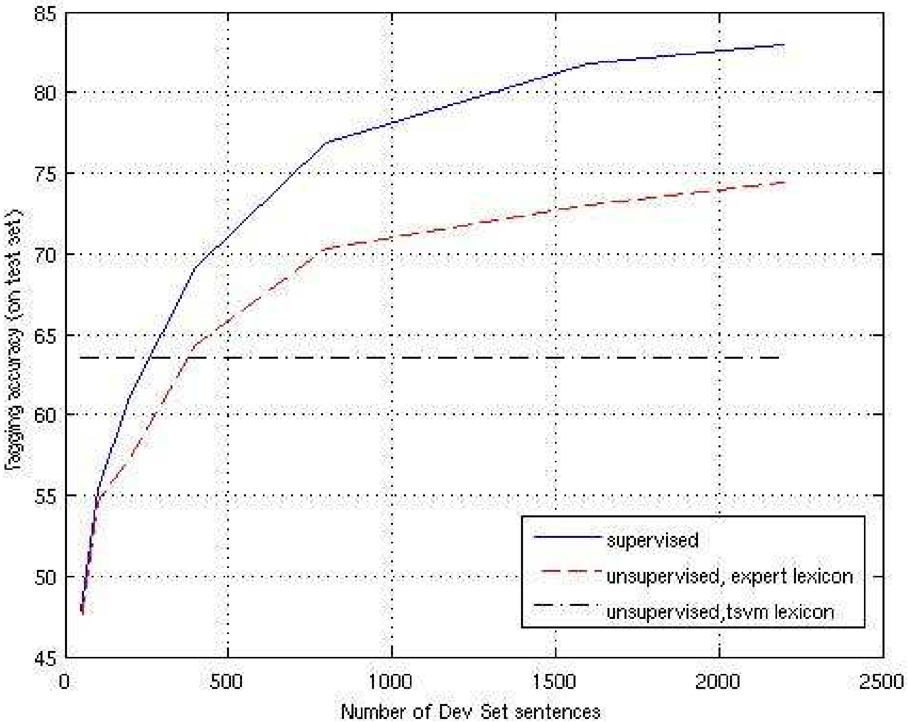
- Fix errors from (Step 1) Morphological analysis
  - Use “oracle” labels collected from Dev Set
  - 1500 of labeled words occur in Dev Set
  - Repair 1000 words

Repair training data	Repair lexicon	Tag Acc. (TSVM)
Y	Y	**66.5**
N	Y	**66.7**
Y	N	64.9
N	N	63.5



Comparisons: when more resources are available

- Unsupervised training, full expert lexicon
  - Collect “oracle” lexicon from Dev Set
- Supervised training (on Dev Set)



NOTE:
- TSVM results use Train Set, not Dev Set