ML for Coreference Resolution

- noun phrase coreference resolution
  - quick review
- a (supervised) machine learning approach
  - the truth this time
- weakly supervised approaches

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> **Queen Elizabeth** set about transforming **her** husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming **her husband**, **King George VI**, into a viable monarch. Logue, a renowned speech therapist, was summoned to help **the King** overcome **his** speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. **Logue, a renowned speech therapist**, was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into <u>a viable monarch</u>. Logue, a renowned speech therapist, was summoned to help the King overcome his <u>speech impediment</u>...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband, King George VI, into <u>a viable monarch</u>. Logue, a renowned speech therapist, was summoned to help the King overcome his <u>speech impediment</u>...

A Machine Learning Approach

- Classification
  - given a description of two noun phrases, NP_i and NP_j, classify the pair as *coreferent* or *not coreferent*

coref ? *coref ?*

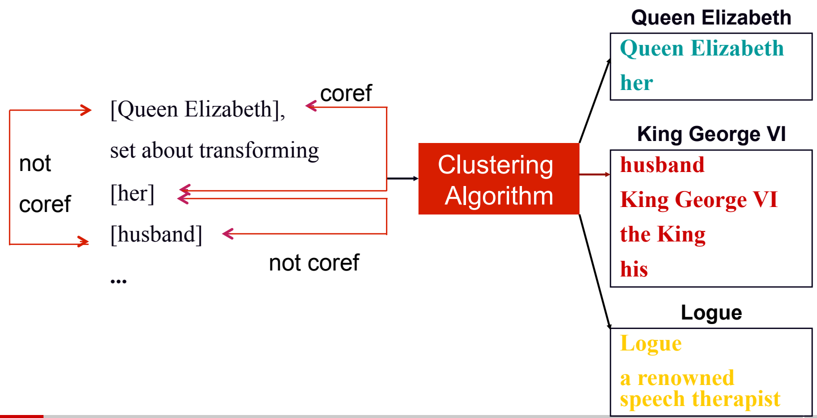
[Queen Elizabeth] set about transforming [her] [husband], ...

coref ?

Aone & Bennett [1995]; Connolly et al. [1994]; McCarthy & Lehnert [1995]; Soon et al. [2001]; Ng & Cardie [2002]; …

A Machine Learning Approach

- Clustering
  - coordinates pairwise coreference decisions



Training Data Creation

- Creating training instances
  - texts annotated with coreference information

candidate antecedent   anaphor

  - one instance *inst*(NP_i, NP_j) for each *ordered* pair of NPs
    - » NP_i precedes NP_j
    - » feature vector: describes the two NPs and context
    - » class value:

coref — pairs on the same coreference chain
not coref — otherwise

Instance Representation

- 25 features per instance
  - lexical (3)
    - » string matching for pronouns, proper names, common nouns
  - grammatical (18)
    - » pronoun_1, pronoun_2, demonstrative_2, indefinite_2, …
    - » number, gender, animacy
    - » appositive, predicate nominative
    - » binding constraints, simple contra-indexing constraints, …
    - » span, maximalnp, …
  - semantic (2)
    - » same WordNet class
    - » alias
  - positional (1)
    - » distance between the NPs in terms of # of sentences
  - knowledge-based (1)
    - » naïve pronoun resolution algorithm

Learning Algorithm

- RIPPER (Cohen, 1995)
  C4.5 (Quinlan, 1994)
  - rule learners
    - » input: set of training instances
    - » output: coreference classifier

- Learned classifier
    - » input: test instance (represents pair of NPs)
    - » output: classification
      confidence of classification

Clustering Algorithm

- Best-first single-link clustering
  - Mark each NP_j as belonging to its own class: $NP_j \in c_j$
  - Proceed through the NPs in left-to-right order.
    - » For each NP, NP_j, create test instances, $inst(NP_i, NP_j)$, for all of its preceding NPs, NP_i.
    - » Select as the antecedent for NP_j the highest-confidence coreferent NP, NP_i, according to the coreference classifier (or none if all have below .5 confidence); Merge c_j and c_j .

Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Ng & Cardie	63.3	76.9	**69.5**	54.2	76.3	**63.4**
Best MUC System	59	72	**65**	56.1	68.8	**61.8**

Baseline Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	40.7	73.5	**52.4**	27.2	86.3	**41.3**
Worst MUC System	36	44	40	52.5	21.4	30.4
Best MUC System	59	72	65	56.1	68.8	61.8
Ng & Cardie	63.3	76.9	**69.5**	54.2	76.3	**63.4**

Detailed Results

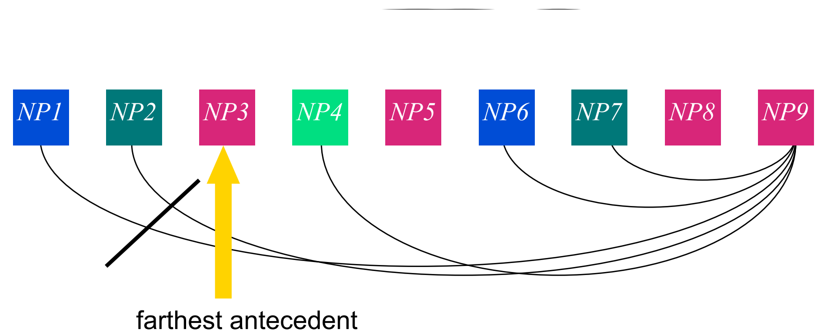
	C4.5						RIPPER					
	MUC-6			MUC-7			MUC-6			MUC-7		
	R	P	F	R	P	F	R	P	F	R	P	F
Original Soon	58.6	67.3	62.6	56.1	65.5	60.4	-	-	-	-	-	-
Duplicated Soon Bsln	64.0	67.0	65.5	55.2	68.5	61.2	62.4	65.0	63.7	54.0	69.5	60.8
Learning Framework	62.4	73.5	67.5	56.3	71.5	63.0	60.8	75.3	67.2	55.3	73.8	**63.2**
All Feats	70.1	58.3	63.6	65.3	56.9	60.8	69.1	62.5	65.6	64.0	55.6	59.5
Hand Feats	64.1	74.9	**69.1**	57.4	70.8	**63.4**	64.2	78.0	**70.4**	55.7	72.8	**63.1**
pronouns	-	77.5	-	-	57.4	-	-	77.9	-	-	56.5	-
proper	-	94.8	-	-	86.6	-	-	94.6	-	-	61.9	-
generic	-	54.7	-	-	64.8	-	-	54.3	-	-	59.9	-

Classifier for MUC-6 Data Set

```
ALIAS = C: +
ALIAS = I:
| SOON_STR_NONPRO = C:
| | ANIMACY = NA: -
| | ANIMACY = I: -
| | ANIMACY = C: +
| SOON_STR_NONPRO = I:
| | PRO_STR = C: +
| | PRO_STR = I:
| | | PRO_RESOLVE = C:
| | | | EMBEDDED_1 = Y: -
| | | | EMBEDDED_1 = N:
| | | | | PRONOUN_1 = Y:
| | | | | | ANIMACY = NA: -
| | | | | | ANIMACY = I: -
| | | | | | ANIMACY = C: +
| | | | | PRONOUN_1 = N:
| | | | | | MAXIMALNP = C: +
| | | | | | MAXIMALNP = I:
| | | | | | | WNCLASS = NA: -
| | | | | | | WNCLASS = I: +
| | | | | | | WNCLASS = C: +
| | | PRO_RESOLVE = I:
| | | | APPOSITIVE = I: -
| | | | APPOSITIVE = C:
| | | | | GENDER = NA: +
| | | | | GENDER = I: +
| | | | | GENDER = C: -
```


Problem 1

- Coreference is a rare relation
  - skewed class distributions (2% positive instances)
  - *remove some negative instances*



Problem 2

- Coreference is a discourse-level problem with different solutions for different types of NPs
  - » proper names: string matching and aliasing
  - inclusion of "hard" positive training instances
  - *positive example selection*: selects easy positive training instances (cf. Harabagiu *et al.* (2001))

Queen Elizabeth set about transforming **her husband**,
King George VI, into a viable monarch. Logue,
the renowned speech therapist, was summoned to help
the King overcome his speech impediment...

Problem 3

- Coreference is an equivalence relation
  - loss of transitivity
  - need to tighten the connection between classification and clustering
  - *prune learned rules w.r.t. the clustering-level coreference scoring function*

Results

	MUC-6 R	MUC-6 P	MUC-6 F	MUC-7 R	MUC-7 P	MUC-7 F
Baseline	40.7	73.5	52.4	27.2	86.3	41.3
NEG-SELECT	46.5	67.8	55.2	37.4	59.7	46.0
POS-SELECT	53.1	80.8	64.1	41.1	78.0	53.8
NEG-SELECT + POS-SELECT	63.4	76.3	69.3	59.5	55.1	57.2
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	69.5	54.2	76.3	63.4

- Ultimately: large increase in F-measure, due to gains in recall

CORNELL

Comparison with Best MUC Systems

	MUC-6 R	MUC-6 P	MUC-6 F	MUC-7 R	MUC-7 P	MUC-7 F
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	69.5	54.2	76.3	63.4
Best MUC System	59	72	65	56.1	68.8	61.8

CORNELL

Supervised ML for NP Coreference

- Good performance compared to other systems, but…**lots** of room for improvement
  - Common nouns < pronouns < proper nouns
  - Tighter connection between classification and clustering is possible
    - » Rich Caruana's (2004) ensemble methods
    - » Statistical methods for learning probabilistic relational models (Getoor *et al.,* 2001; Lafferty et al., 2001; Taskar *et al.*, 2003; McCallum and Wellner, 2003).
  - Need additional data sets
    - » ACE data from Penn's LDC
    - » General problem: reliance on manually annotated data…

CORNELL

Plan for the Talk

- noun phrase coreference resolution
- a (supervised) machine learning approach
- ➡ weakly supervised approaches
  - background
  - two techniques
  - evaluation

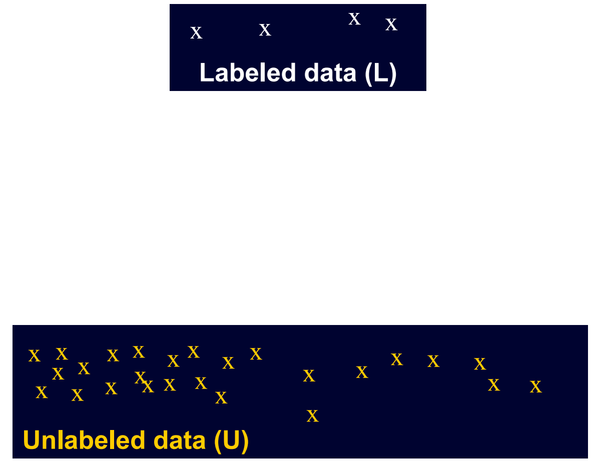
CORNELL

Weakly Supervised Approaches

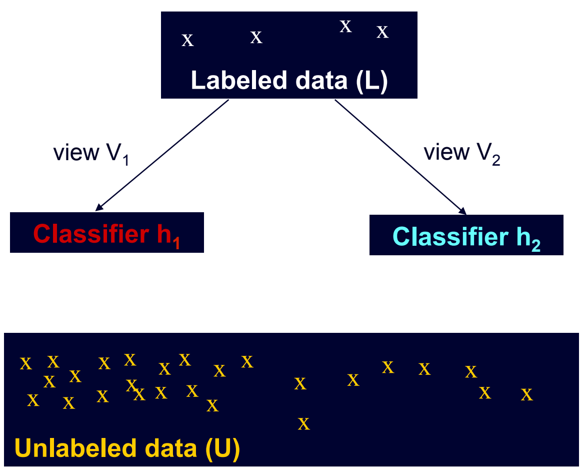
- Idea:
  bootstrap (NP coreference) classifiers using a *small amount of labeled data* (expensive) and a *large amount of unlabeled data* (cheap)

- Methods
  - Co-training
  - Self-training

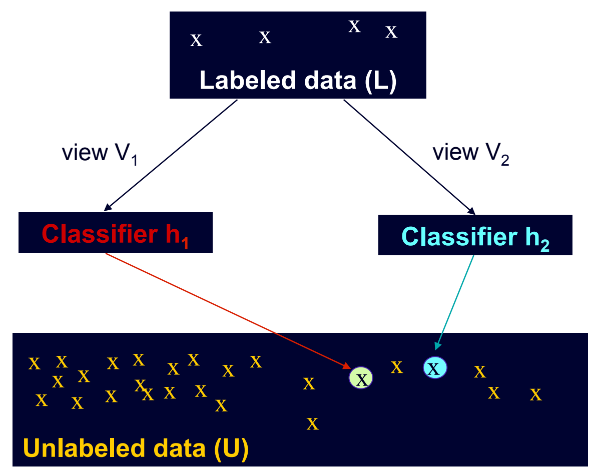
Co-Training [Blum and Mitchell, 1998]



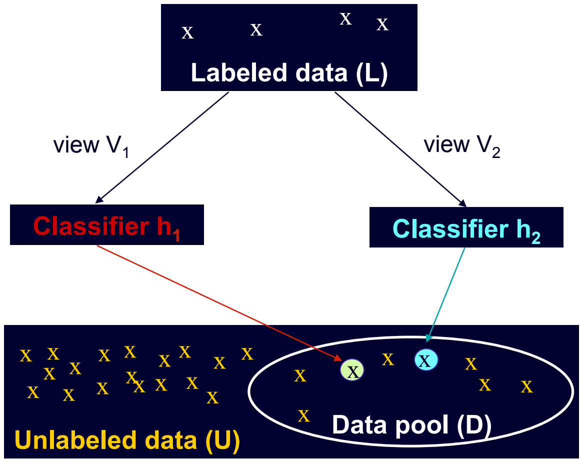
Co-Training [Blum and Mitchell, 1998]



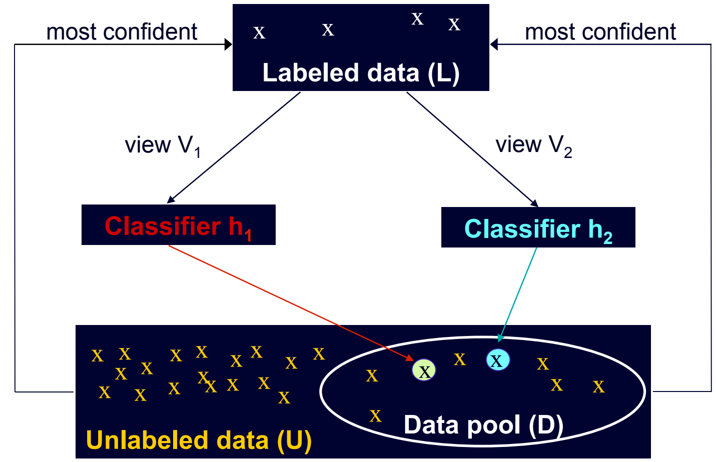
Co-Training [Blum and Mitchell, 1998]



Co-Training [Blum and Mitchell, 1998]



Co-Training [Blum and Mitchell, 1998]



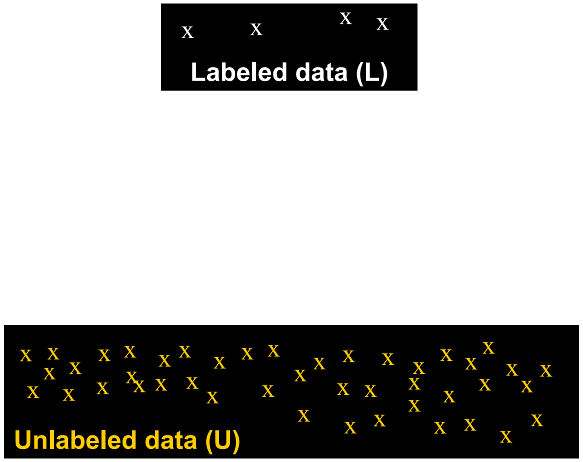
Potential Problems with Co-Training

- Strong assumptions on the views (Blum and Mitchell, 1998)
  - each view must be sufficient for learning the target concept
  - the views must be conditionally independent given the class
  - empirically shown to be sensitive to these assumptions (Muslea *et al*., 2002)
- A number of parameters need to be tuned
  - views, data pool size, growth size, number of iterations, initial size of labeled data
  - algorithm is sensitive to its input parameters (Nigam and Ghani, 2000; Pierce and Cardie, 2001; Pierce 2003)

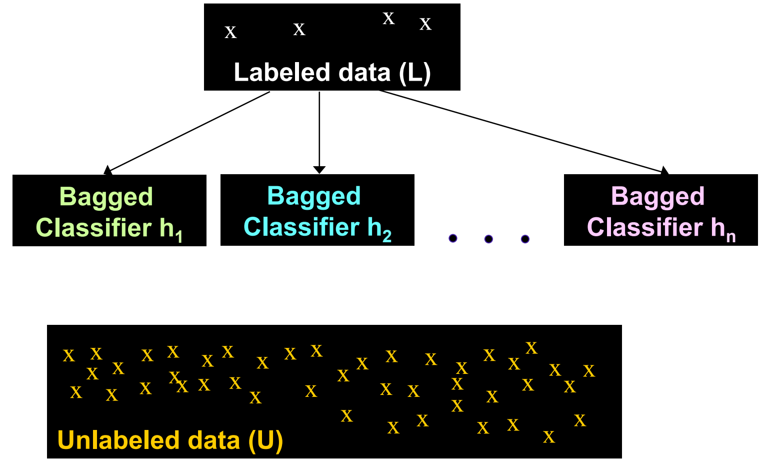
Potential Problems with Co-Training

- Multi-view algorithm
  - Is there any natural feature split for NP coreference?
    - » view factorization is a non-trivial problem for coreference
      - ◆ Mueller *et al*.'s (2002) greedy method

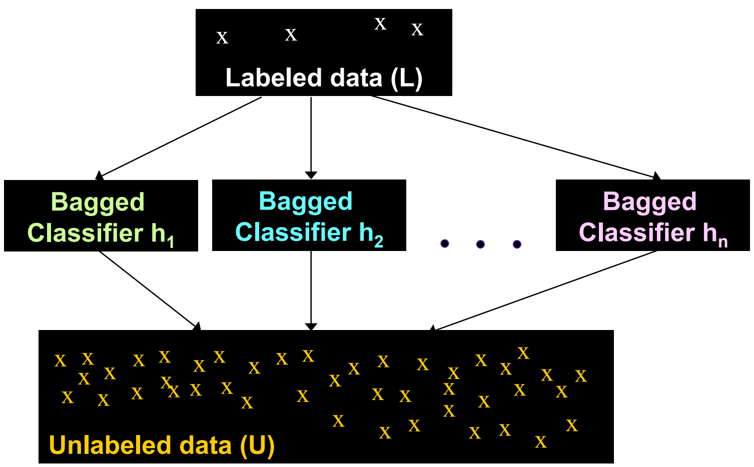
Self-Training with Bagging
[Banko and Brill, 2001]



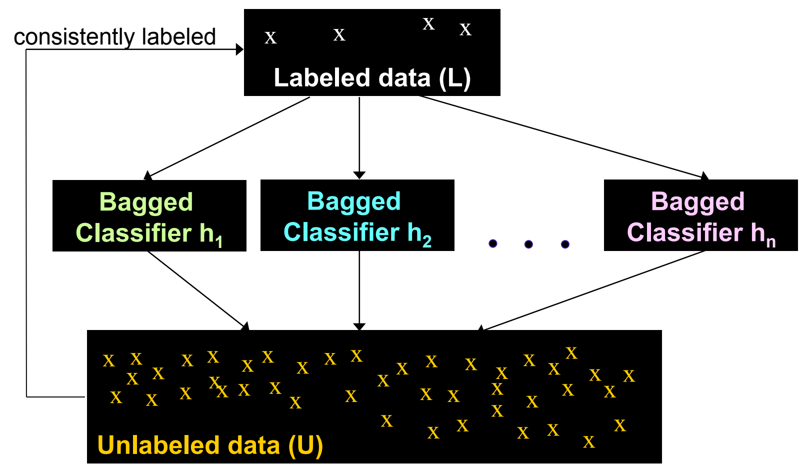
Self-Training with Bagging
[Banko and Brill, 2001]



Self-Training with Bagging
[Banko and Brill, 2001]



Self-Training with Bagging
[Banko and Brill, 2001]



Plan for the Talk

- noun phrase coreference resolution
- a (supervised) machine learning approach
- weakly supervised approaches
  - background
  - two techniques
  - evaluation

CORNELL

Evaluation

- MUC-6 and MUC-7 coreference data sets
- labeled data (L): one dryrun text
  - » 3500-3700 instances
- unlabeled data (U): remaining 29 dryrun texts
- vs. fully supervised ML
  - ~500,000 instances (30 dryrun texts)

CORNELL

Results (Baseline)

- train a naïve Bayes classifier on the single (labeled) text using all 25 features

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	58.3	52.9	**55.5**	52.8	37.4	**43.8**

CORNELL

Evaluating the Weakly Supervised Algorithms

- Determine the best parameter setting of each algorithm (in terms of its effectiveness in improving performance)

CORNELL

Co-Training Parameters

- Views (3 heuristic methods for view factorization)
  - Mueller *et al.*'s (2002) greedy method
  - random splitting
  - splitting according to the feature type
- Pool size
  - 500, 1000, 5000
- Growth size
  - 10, 50, 100, 200, 250
- Number of co-training iterations
  - run until performance stabilized

Results (Co-Training)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	58.3	52.9	**55.5**	52.8	37.4	**43.8**
Co-Training	47.5	81.9	**60.1**	40.6	77.6	**53.3**

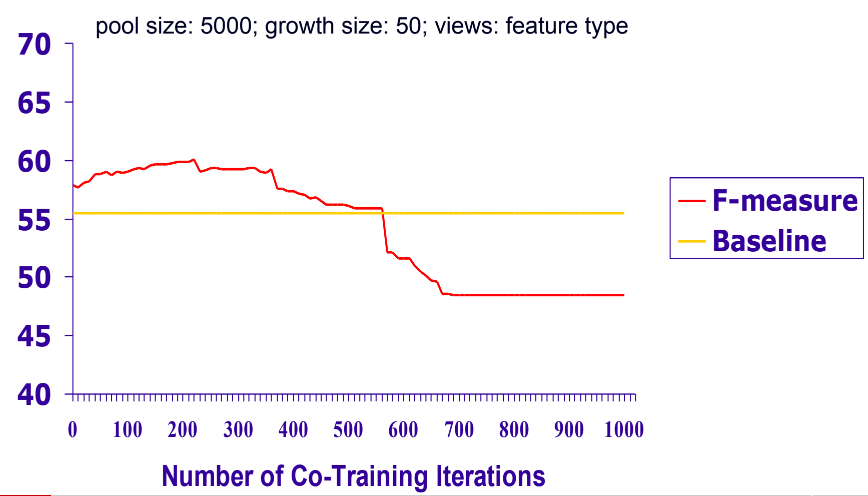
- co-training produces improvements over the baseline at its best parameter settings

Results (Co-Training)

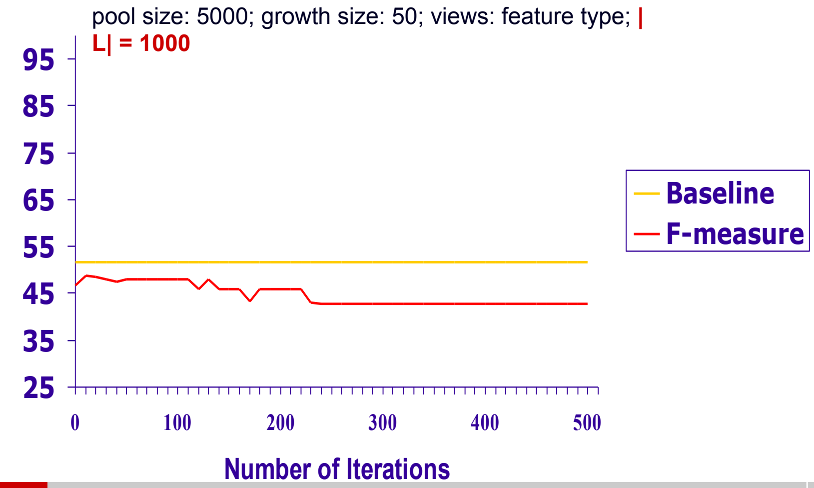
	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	58.3	52.9	**55.5**	52.8	37.4	**43.8**
Co-Training	47.5	81.9	**60.1**	40.6	77.6	**53.3**
Supervised ML* (~500,000 insts)	63.3	76.9	69.5	54.2	76.3	63.4

- co-training produces improvements over the baseline at its best parameter settings

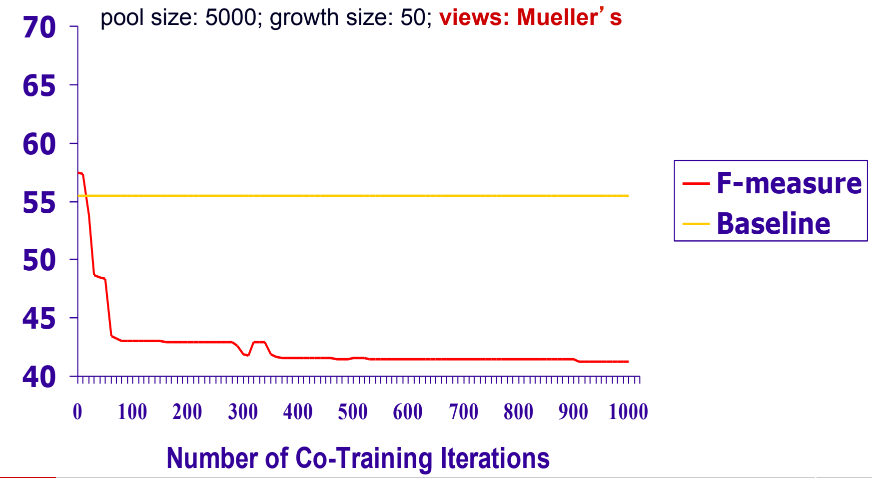
Learning Curve for Co-Training (MUC-6)



Learning Curve for Co-Training (MUC-6)



Learning Curve for Co-Training (MUC-6)



Self-Training Parameters

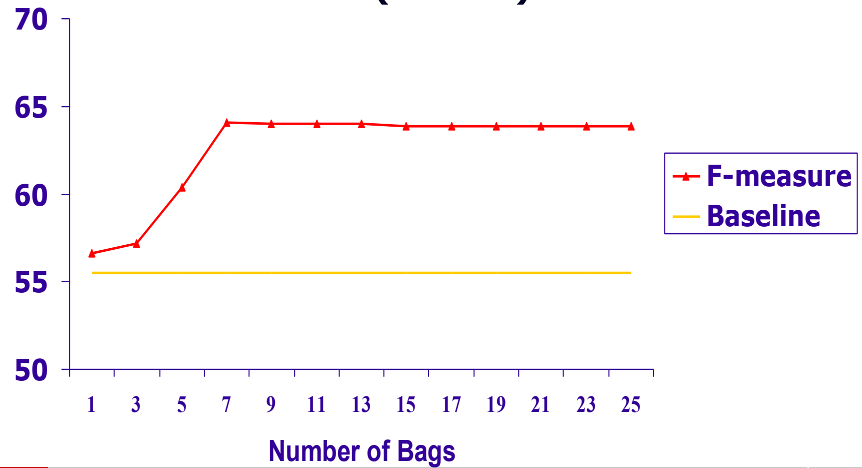
- Number of bags
  - tested all odd number of bags between 1 and 25
- 25 bags are sufficient for most learning tasks (Breiman, 1996)

Results (Self-Training with Bagging)

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	58.3	52.9	55.5	52.8	37.4	43.8
Co-Training	47.5	81.9	60.1	40.6	77.6	53.3
Self-Training with Bagging	54.1	78.6	64.1	54.6	62.6	58.3

- Self-training performs better than co-training

Self-Training: Effect of the Number of Bags (MUC-6)



Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	58.3	52.9	**55.5**	52.8	37.4	**43.8**
Co-Training	47.5	81.9	**60.1**	40.6	77.6	**53.3**
Self-Training with Bagging	54.1	78.6	**64.1**	54.6	62.6	**58.3**
Supervised ML* (~500,000 insts)	63.3	76.9	69.5	54.2	76.3	63.4

Summary

- Supervised ML approach to NP coreference resolution
  - Good performance relative to other approaches
  - Still lots of room for improvement
- Weakly supervised approaches are promising
  - Not as good performance as fully supervised, but use much less manually annotated training data
- For problems where no natural view factorization exists…
  - Single-view weakly supervised algorithms
    - » Self-training with bagging