Practical Active Learning for Complex NLP Tasks

Florian Laws

08.08.2012

Motivation

Supervised classification

- successful for many NLP tasks.

But. . .

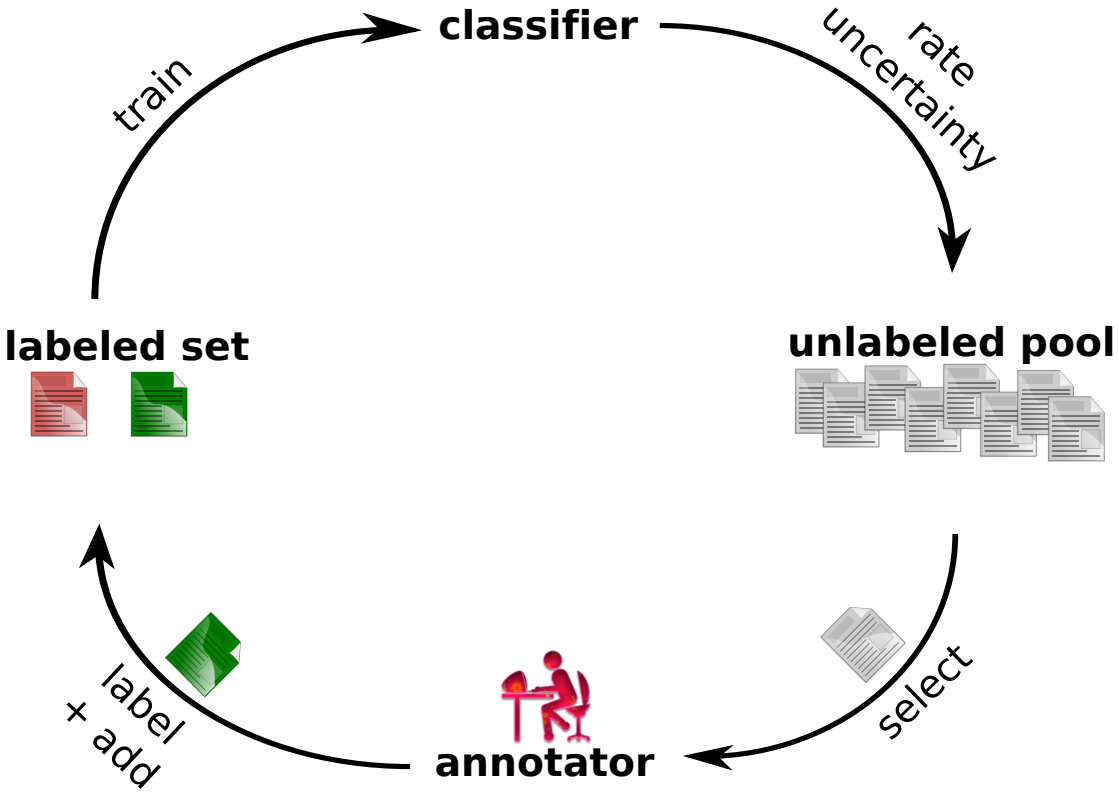
- needs lots of labeled training data.
- labeled data is tedious and expensive to create.

- Use fewer manually annotated examples.

$\Rightarrow$ Active Learning (AL)

Active Learning (AL)

Interactive annotation loop:



Use only informative examples!

Outline

Contributions

Robust and efficient training example selection for NER and coreference resolution.

Stop annotation without wasted annotation effort.

Obtain low-cost AL annotations on the web.

Outline

Contributions

Robust and efficient training example selection for NER and coreference resolution.

Stop annotation without wasted annotation effort.

Obtain low-cost AL annotations on the web.

Outline

Contributions

Robust and efficient training example selection for NER and coreference resolution.

Stop annotation without wasted annotation effort.

Obtain low-cost AL annotations on the web.

Outline

Contributions

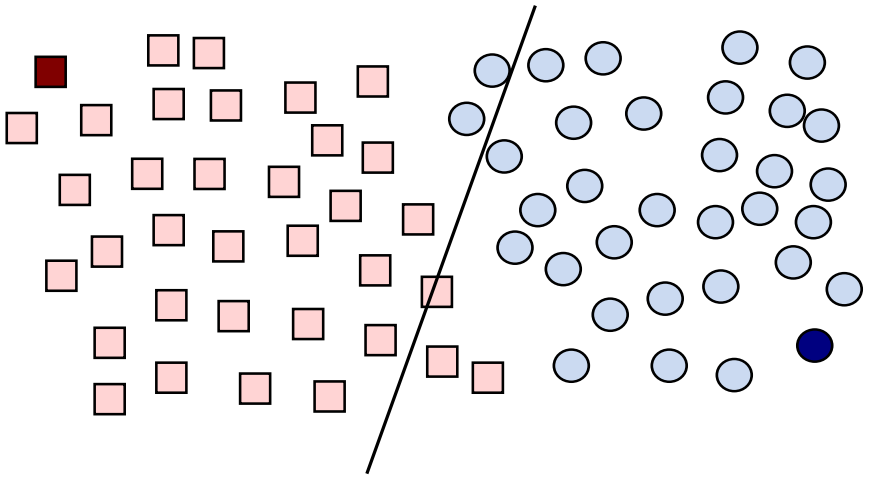
Robust and efficient training example selection for NER and coreference resolution.

Stop annotation without wasted annotation effort.

Obtain low-cost AL annotations on the web.

Random Selection vs. Active Learning

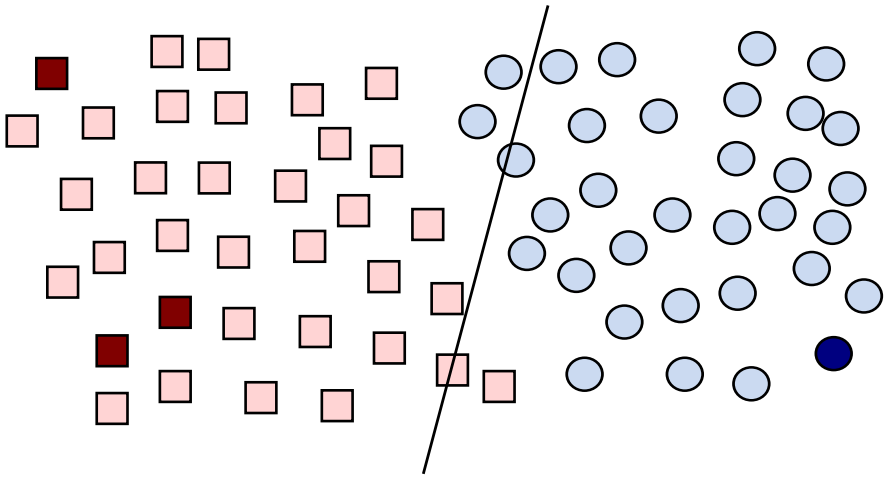
Random Selection



Active Learning

Random Selection vs. Active Learning

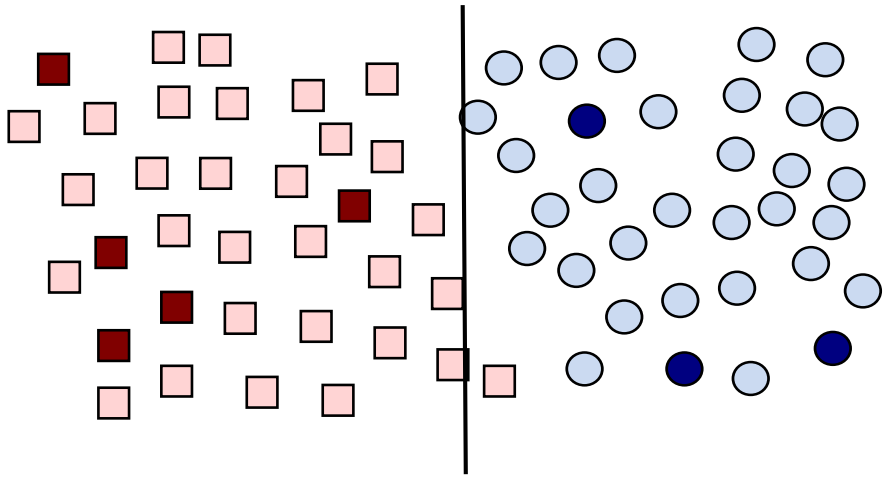
Random Selection



Active Learning

Random Selection vs. Active Learning

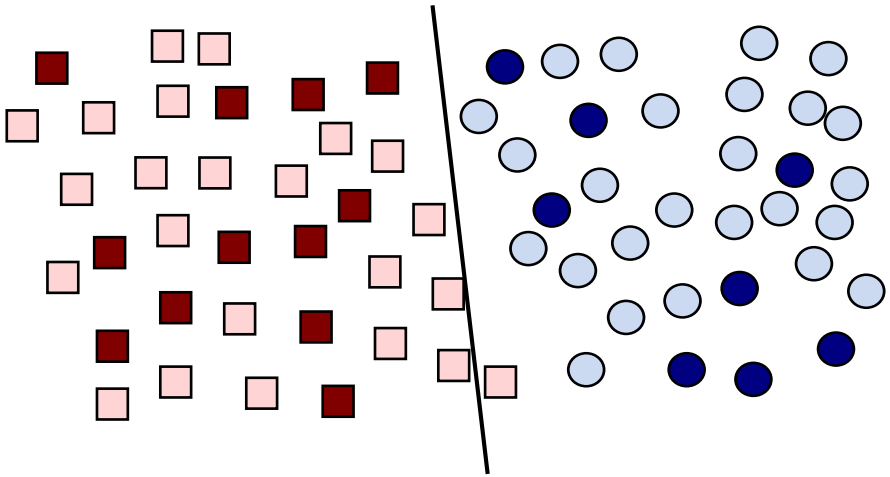
Random Selection



Active Learning

Random Selection vs. Active Learning

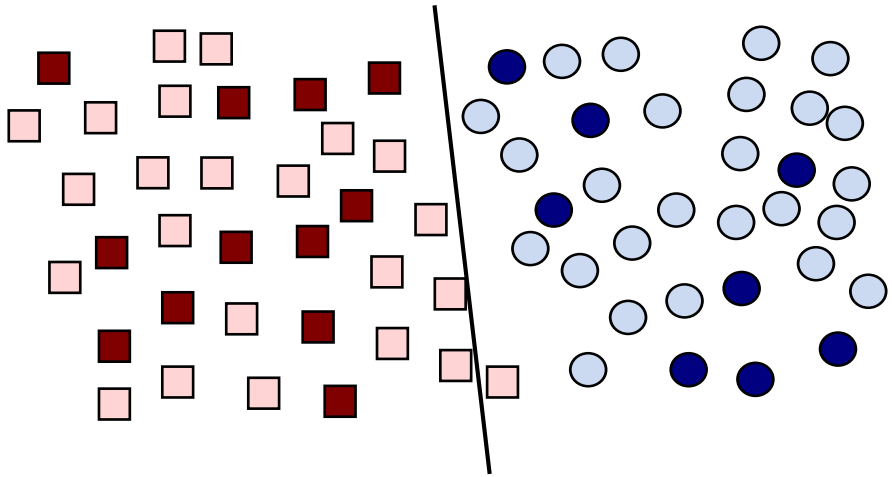
Random Selection



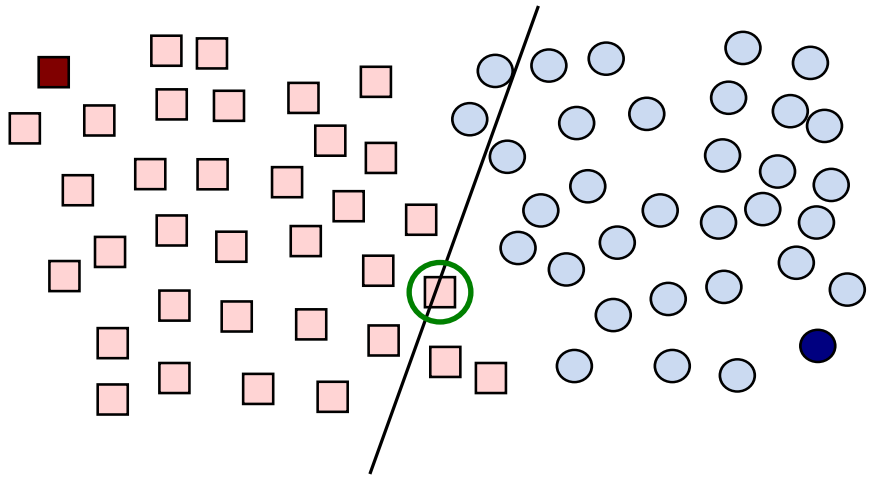
Active Learning

Random Selection vs. Active Learning

Random Selection

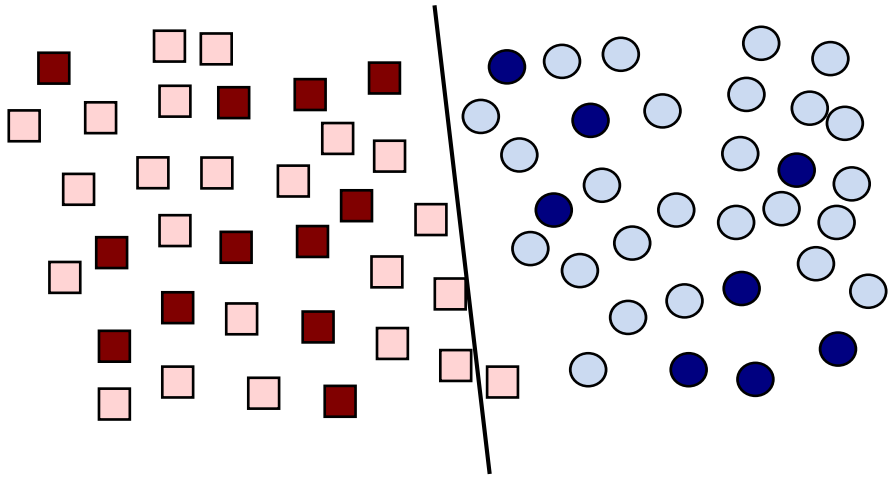


Active Learning

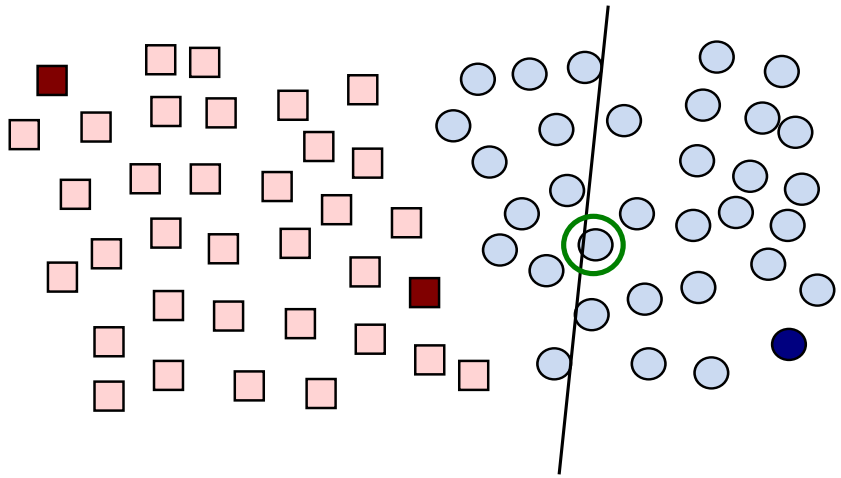


Random Selection vs. Active Learning

Random Selection

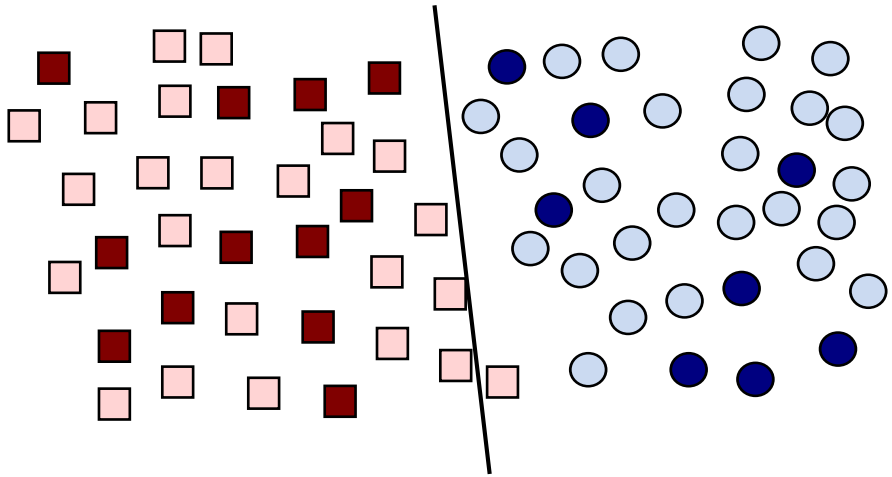


Active Learning

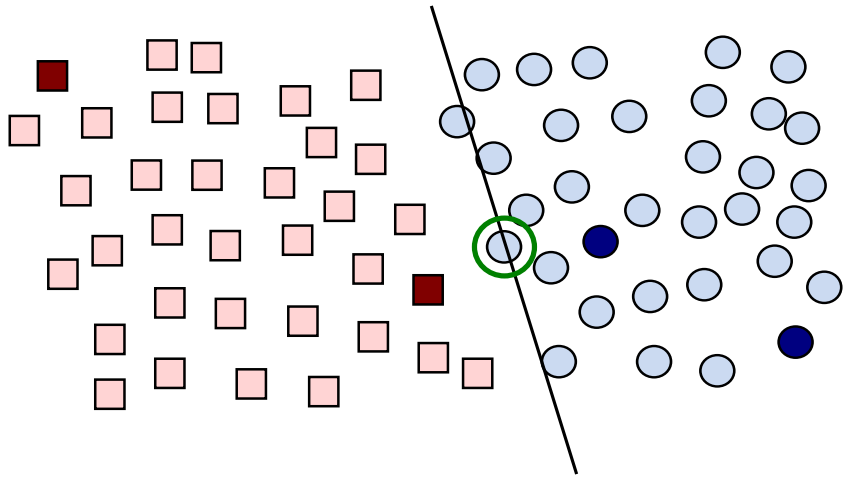


Random Selection vs. Active Learning

Random Selection

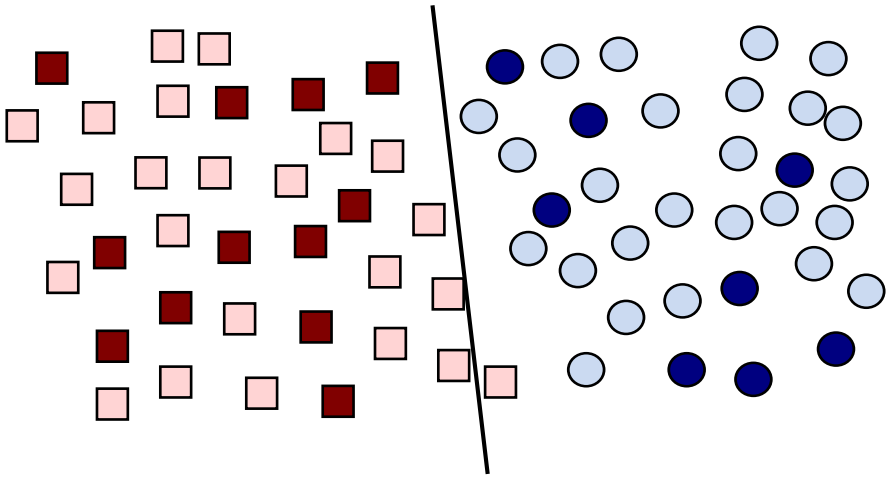


Active Learning

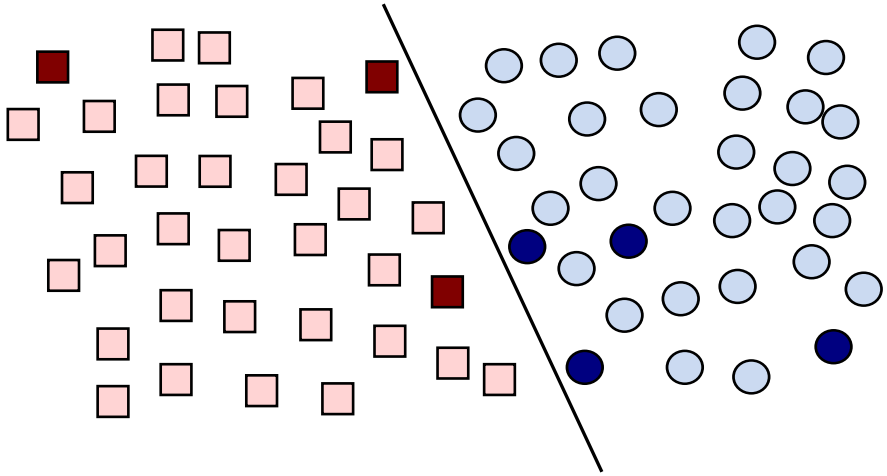


Random Selection vs. Active Learning

Random Selection



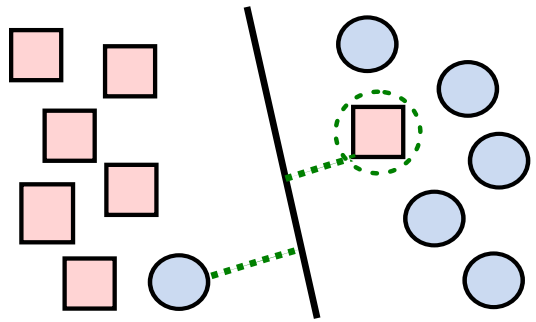
Active Learning



How to select informative examples?

Uncertainty Sampling

- "Distance" to decision boundary.
- Based on class probabilities.
- Choice of measures for multiclass classification: (Entropy, Margin, etc.).
- Simple, fast.
- Used in NER experiments.



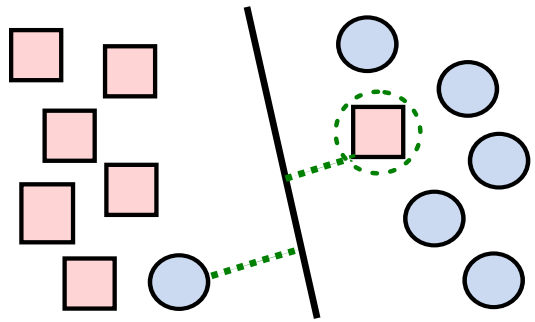
Query by Committee

- Train multiple classifiers.
- Select examples where classifiers disagree.
- Usable with any classifier.
- Basis for Coref experiments.

How to select informative examples?

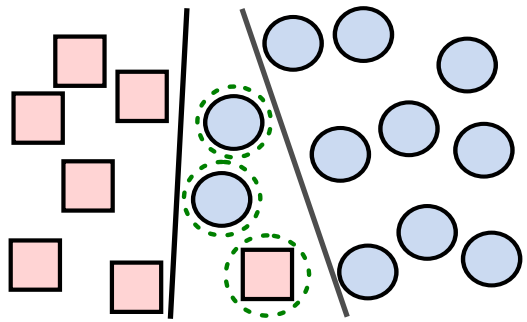
Uncertainty Sampling

- "Distance" to decision boundary.
- Based on class probabilities.
- Choice of measures for multiclass classification: (Entropy, Margin, etc.).
- Simple, fast.
- Used in NER experiments.



Query by Committee

- Train multiple classifiers.
- Select examples where classifiers disagree.
- Usable with any classifier.
- Basis for Coref experiments.

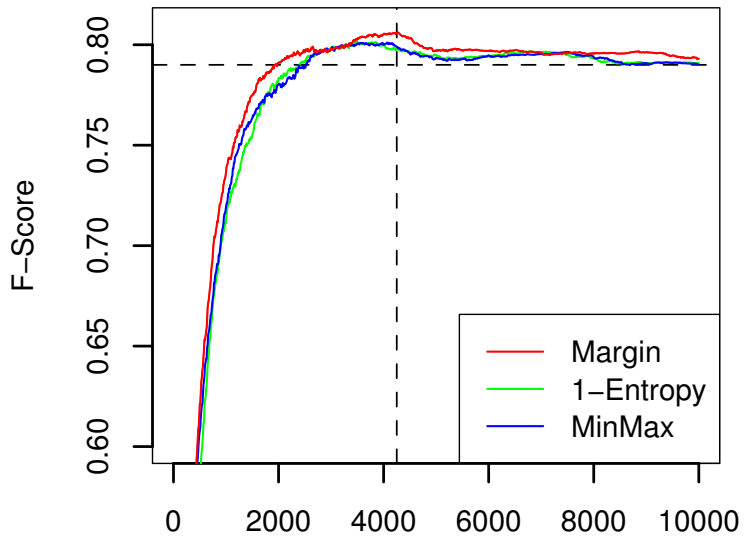


Outline

Active Learning for Named Entity Recognition

NER as Tagging:

" Gordon^{PER} Brown^{PER} is^{O} arguing^{O} that^{O} a^{O} WTO^{ORG} . . . "

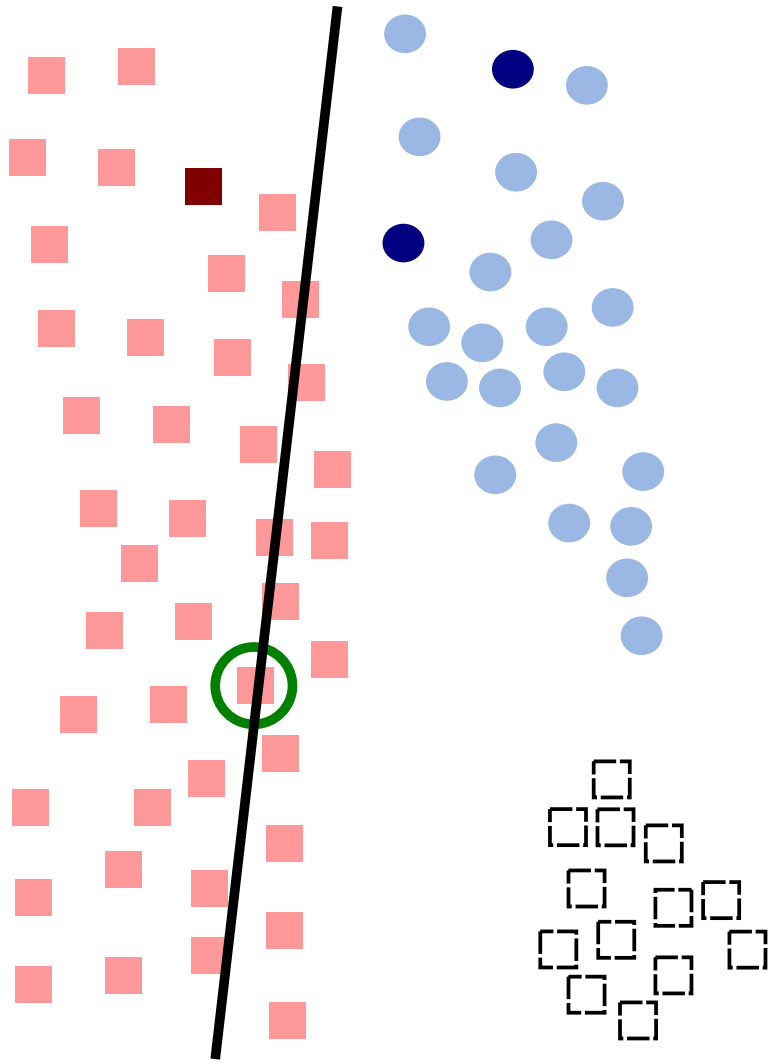


- 12% of examples needed to reach supervised baseline.
- Peak at 20% of examples.
- Margin best uncertainty measure (slightly).

The Missed Cluster/Missed Class problem

- AL can miss parts of the sample space!
- Slow learning on some classes: Missed cluster/class problem.

- Exploration of sample space can give better solution.
  - Some annotation effort is spent on “uninformative” examples.
  - Which ones to choose?
- *Co-selection* can provide some exploration at low effort.

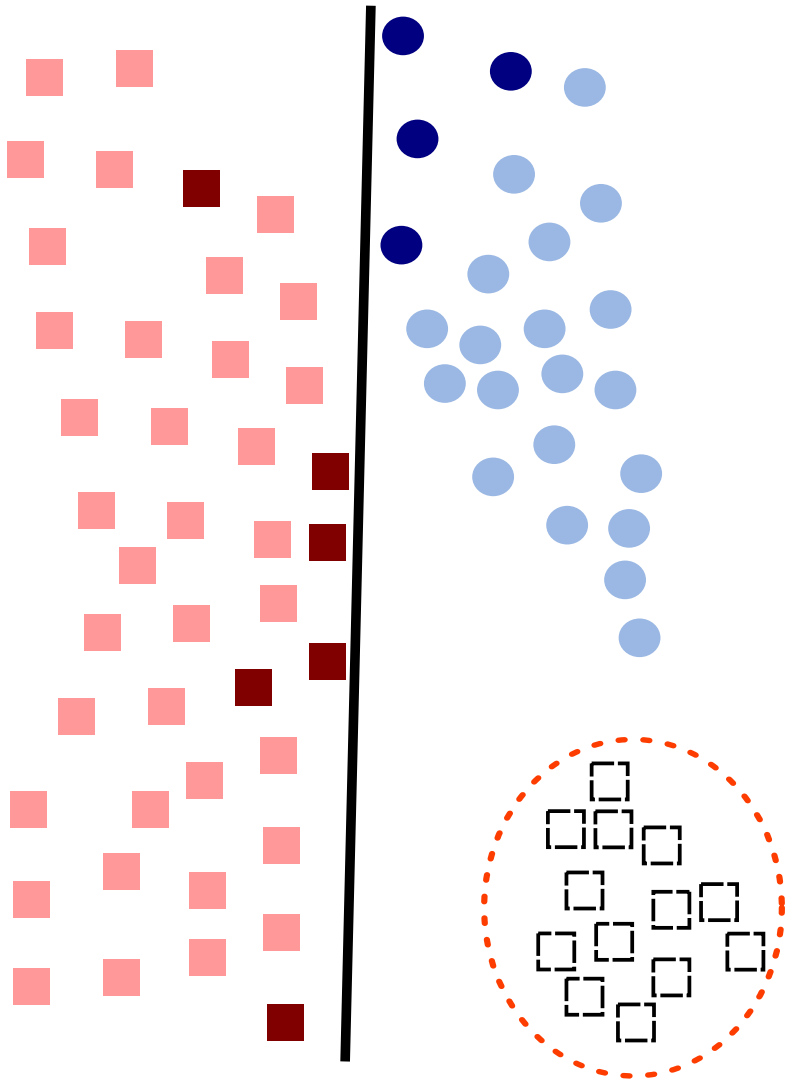


(joint work with Katrin Tomanek.)

The Missed Cluster/Missed Class problem

- AL can miss parts of the sample space!
- Slow learning on some classes: Missed cluster/class problem.

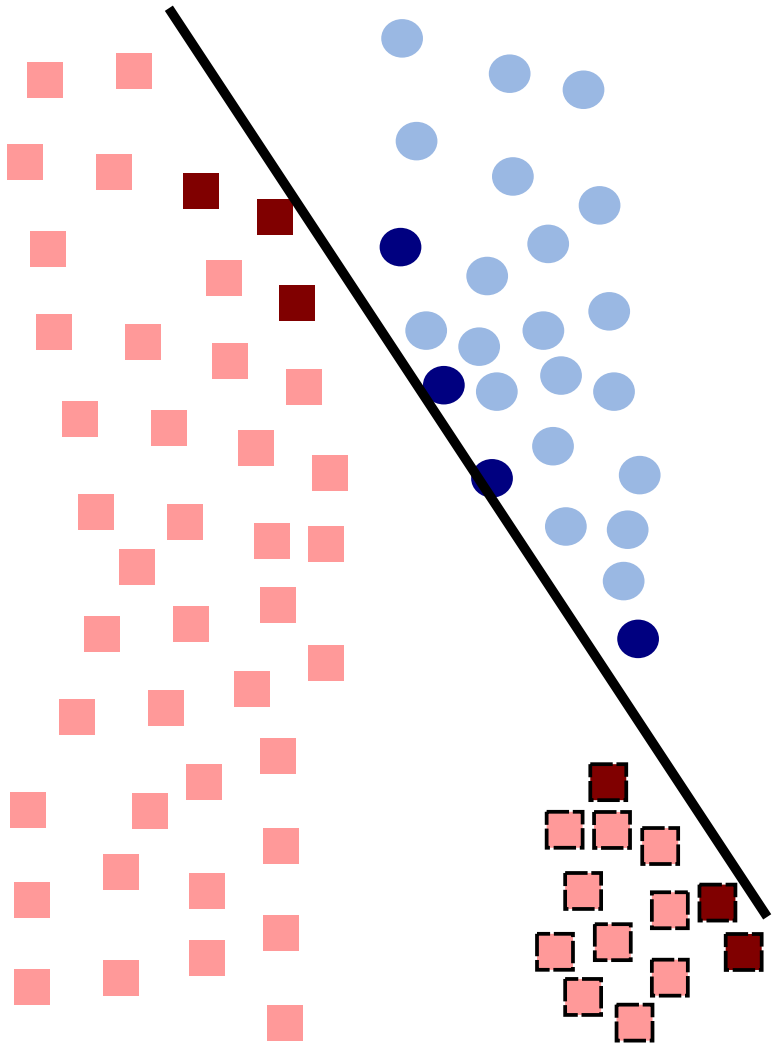
- Exploration of sample space can give better solution.
  - Some annotation effort is spent on “uninformative” examples.
  - Which ones to choose?
- *Co-selection* can provide some exploration at low effort.



(joint work with Katrin Tomanek.)

The Missed Cluster/Missed Class problem

- AL can miss parts of the sample space!
- Slow learning on some classes: Missed cluster/class problem.



- Exploration of sample space can give better solution.
  - Some annotation effort is spent on “uninformative” examples.
  - Which ones to choose?
- *Co-selection* can provide some exploration at low effort.

(joint work with Katrin Tomanek.)

The co-selection effect.

Token selection:

"[Gordon Brown]$^{\text{uncertain}}$ is arguing that a WTO deal on trade barriers in Geneva could be critical in bringing food prices under control."

Sentence selection:

"Gordon [Brown]$^{\text{uncertain}}$ is arguing that a WTO deal on trade barriers in Geneva could be critical in bringing food prices under control."

- **Co-selection:** Get examples for other classes as a side effect!
- Classes must co-occur in sentences
- Co-selection can provide exploration at little extra effort!

The co-selection effect.

Token selection:

"[Gordon Brown]^{PER} is arguing that a WTO deal on trade barriers in Geneva could be critical in bringing food prices under control."

Sentence selection:

"Gordon [Brown]^{uncertain} is arguing that a WTO deal on trade barriers in Geneva could be critical in bringing food prices under control."

- **Co-selection:** Get examples for other classes as a side effect!
- Classes must co-occur in sentences
- Co-selection can provide exploration at little extra effort!

The co-selection effect.

Token selection:

"[Gordon Brown]^{PER} is arguing that a WTO deal on trade barriers in Geneva could be critical in bringing food prices under control."

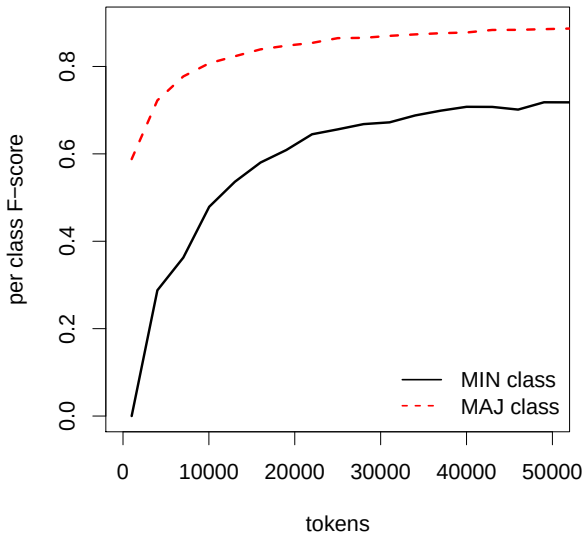
Sentence selection:

"[Gordon Brown]^{PER} is arguing that a [WTO]^{ORG} deal on trade barriers in [Geneva]^{GPE} could be critical in bringing food prices under control."

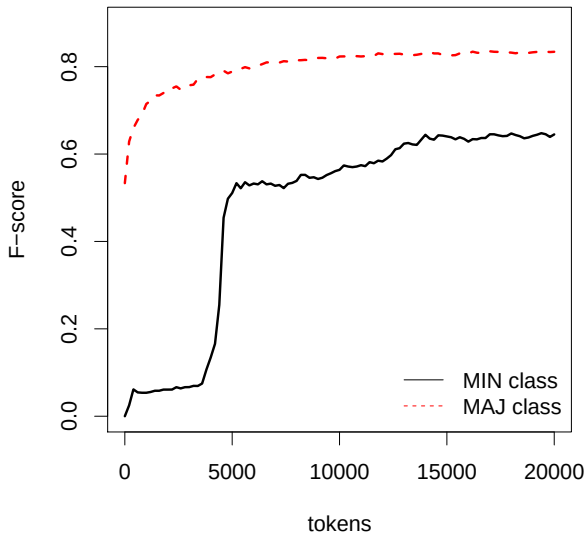
- **Co-selection:** Get examples for other classes as a side effect!
- Classes must co-occur in sentences
- Co-selection can provide exploration at little extra effort!

Co-selection helps to make learning more robust.

NER on Penn BioIE dataset, no entities in seed set:



Sentence selection



Token selection

Co-selection by sentence selection:

- is more robust.
- may be beneficial from a user interface perspective.

Coreference resolution

Identify mentions of the same real-world entity.

Example:

Rolls-Royce Motor Cars Inc. said it expects its U.S. sales remain steady at about 1,200 cars in 1990. The luxury auto maker last year sold 1,214 cars in the U.S. Howard Mosher, president and chief executive officer, said he anticipates growth for the luxury automaker in Britain and Europe.

Coreference chain (cluster):

"Rolls-Royce" – "it" – "The luxury auto maker" – "the luxury auto maker"

Coreference link (pair):

e.g. "Rolls-Royce" – "it" "Howard Mosher" – "he"

Coreference Resolution and Active Learning

Mention-Pair-Coreference System

1. Link classification: For each pair of markables, predict if coreferent.
2. Combine coreferent links into coreference chains.

Active Learning for Link Classification stage?

- Link Classifier is statistical classifier.
- Can we use Active Learning for training?
- Prior work has failed to show that this works for coref!

Coreference Resolution and Active Learning

Mention-Pair-Coreference System

1. Link classification: For each pair of markables, predict if coreferent.
2. Combine coreferent links into coreference chains.

Active Learning for Link Classification stage?

- Link Classifier is statistical classifier.
- Can we use Active Learning for training?
- Prior work has failed to show that this works for coref!

Fixing AL for coreference resolution

Simple uncertainty sampling does not work

- Uncertainty ratings not reliable for Decision Trees or Naive Bayes

$\Rightarrow$ Use Query-by-committee.

Class imbalance

- Only 1.5% of all links are positive examples.

$\Rightarrow$ Use class-balancing and co-selection:

- Neighborhood pooling (class-balancing):
  - Only use links close to a mention.
  - The neighborhood pool is a subset of the pool of all links.
- Neighborhood sampling: Sample neighborhood as a group.
  - Co-Selection.

Fixing AL for coreference resolution

Simple uncertainty sampling does not work

- Uncertainty ratings not reliable for Decision Trees or Naive Bayes

$\Rightarrow$ Use Query-by-committee.

Class imbalance

- Only 1.5% of all links are positive examples.

$\Rightarrow$ Use class-balancing and co-selection:

- Neighborhood pooling (class-balancing):
  - Only use links close to a mention.
  - The neighborhood pool is a subset of the pool of all links.
- Neighborhood sampling: Sample neighborhood as a group.
  - Co-Selection.

Fixing AL for coreference resolution

Simple uncertainty sampling does not work

- Uncertainty ratings not reliable for Decision Trees or Naive Bayes

$\Rightarrow$ Use Query-by-committee.

Class imbalance

- Only 1.5% of all links are positive examples.

$\Rightarrow$ Use class-balancing and co-selection:

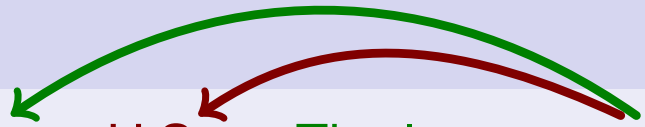
- Neighborhood pooling (class-balancing):
  - Only use links close to a mention.
  - The neighborhood pool is a subset of the pool of all links.
- Neighborhood sampling: Sample neighborhood as a group.
  - Co-Selection.

Neighborhoods

Neighborhood of a target mention:

- Closest positive link to the left of target mention.
- All negative links inbetween.
- Ignore other links.

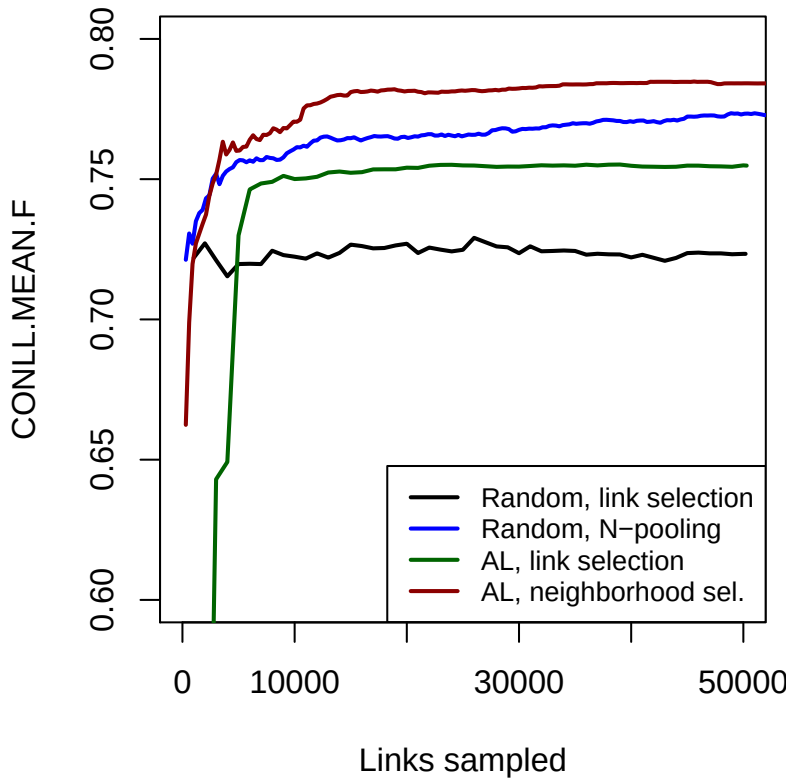
Example



Rolls-Royce Motor Cars Inc. . . . it . . . U.S. . . . The luxury auto maker . . . the U.S. Howard Mosher, president and chief executive officer, . . . he . . . the luxury automaker . . .

- N.-bootstrapping: Guess positive links using the classifiers.
- N.-sampling: Sample entire neighborhoods for annotation.

Experiments on SemEval 2010 dataset



- Neighborhood sampling outperforms link sampling.
- AL with neighborhood sampling outperforms random sampling.
- Successful application of AL to coreference annotation.

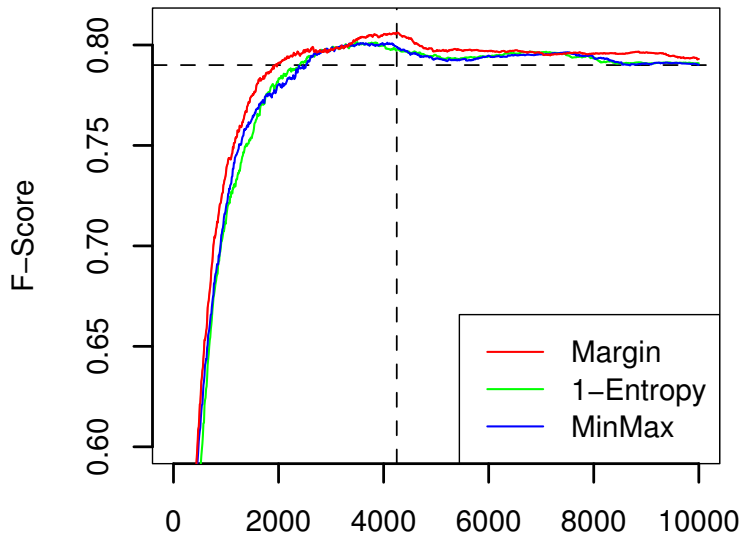
Example selection - Summary

- Active Learning is effective for NER: Needs only 20% of data.
- Co-selection:
  - Enhance robustness by selecting whole sentences.
  - Adresses missed class problem.
- First successful application of AL to Coreference Resolution. using Neighborhood pooling + sampling.

Outline

Monitoring and early stopping are important.

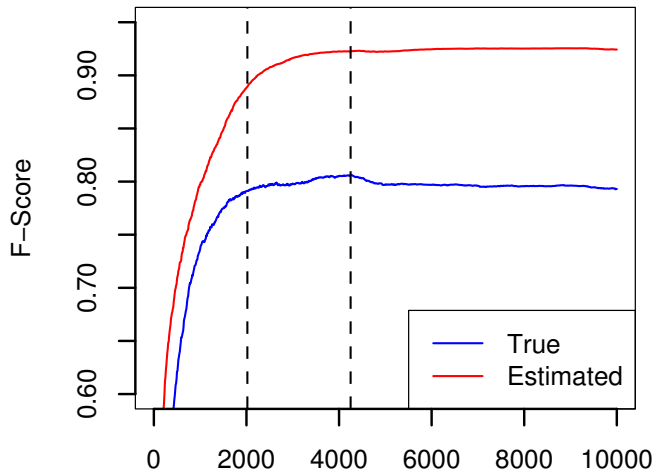
Recap: AL results for NER.



- Peak performance at 20% of pool.
  - Further annotation is wasted effort.
  - Need to stop in time.
- How good is the classifier that is being trained?
  - Performance estimation (F-Score expectation) (see COLING 2008 paper).

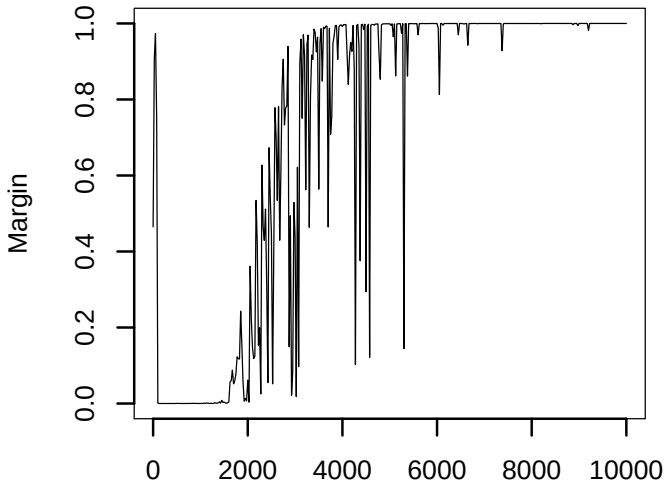
Gradient-based stopping criteria

Performance gradient



- Performance estimate
- overoptimistic, but gradient can be used.
- if Gradient < ε: stop.

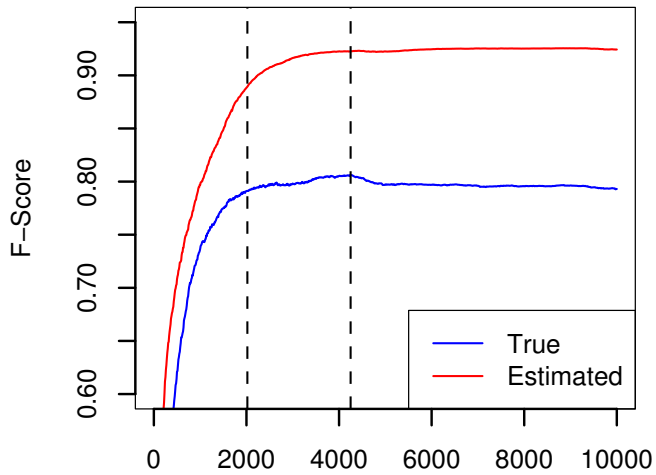
Uncertainty gradient



- Uncertainty of selected example.
- Approaches 1 with training.
- Calculate gradient: if Gradient < ε: stop.
- Filter the noise.

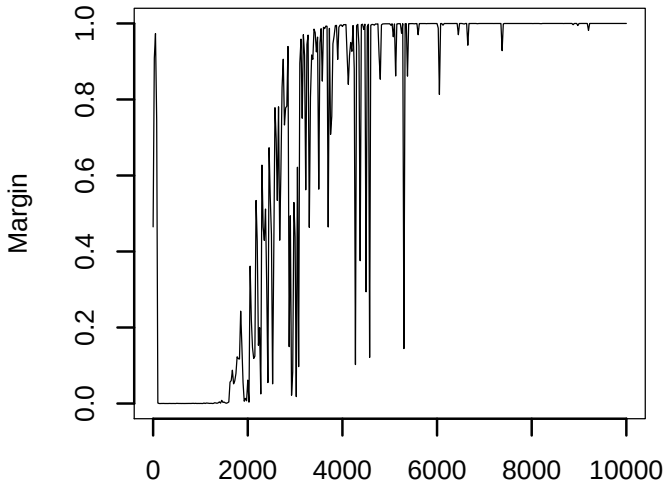
Gradient-based stopping criteria

Performance gradient



- Performance estimate
- overoptimistic, but gradient can be used.
- if Gradient < ε: stop.

Uncertainty gradient



- Uncertainty of selected example.
- Approaches 1 with training.
- Calculate gradient:
  if Gradient < ε: stop.
- Filter the noise.

Results - stopping criteria

NER on ACE 2005, Margin selection, 10-fold validation

Baseline F: 78.7 Peak F: 80.8

Compare to: uncertainty threshold criterion (Zhu and Hovy 2007)

Stop criterion	ε	**Stop at**	Δ Bl	sd	Δ Pk	sd
Uncertainty threshold	0.01	**10.4%**	1.1	0.8	−1.0	0.8
Uncertainty gradient	5e-5	**17.5%**	0.81	0.3	−1.32	0.4
Lewis grad. (Median)	5e-5	**9.2%**	0.8	1.4	−1.3	1.4
Lewis grad. (Mean)	5e-5	**13.1%**	1.1	0.8	−0.95	0.6

- **large reduction in annotation effort**
- stopping point better than baseline
- reliable stopping

Results - stopping criteria

NER on ACE 2005, Margin selection, 10-fold validation

Baseline F: 78.7 Peak F: 80.8

Compare to: uncertainty threshold criterion (Zhu and Hovy 2007)

Stop criterion	ε	Stop at	Δ **BI**	sd	Δ Pk	sd
Uncertainty threshold	0.01	10.4%	**1.1**	0.8	−1.0	0.8
Uncertainty gradient	5e-5	17.5%	**0.81**	0.3	−1.32	0.4
Lewis grad. (Median)	5e-5	9.2%	**0.8**	1.4	−1.3	1.4
Lewis grad. (Mean)	5e-5	13.1%	**1.1**	0.8	−0.95	0.6

- large reduction in annotation effort
- **stopping point better than baseline**
- reliable stopping

Results - stopping criteria

NER on ACE 2005, Margin selection, 10-fold validation

Baseline F: 78.7 Peak F: 80.8

Compare to: uncertainty threshold criterion (Zhu and Hovy 2007)

Stop criterion	ε	Stop at	Δ Bl	sd	Δ Pk	sd
Uncertainty threshold	0.01	10.4%	1.1	0.8	−1.0	0.8
Uncertainty gradient	5e-5	17.5%	0.81	**0.3**	−1.32	0.4
Lewis grad. (Median)	5e-5	9.2%	0.8	1.4	−1.3	1.4
Lewis grad. (Mean)	5e-5	13.1%	1.1	0.8	−0.95	0.6

- large reduction in annotation effort
- stopping point better than baseline
- **reliable stopping**

Monitoring - Summary

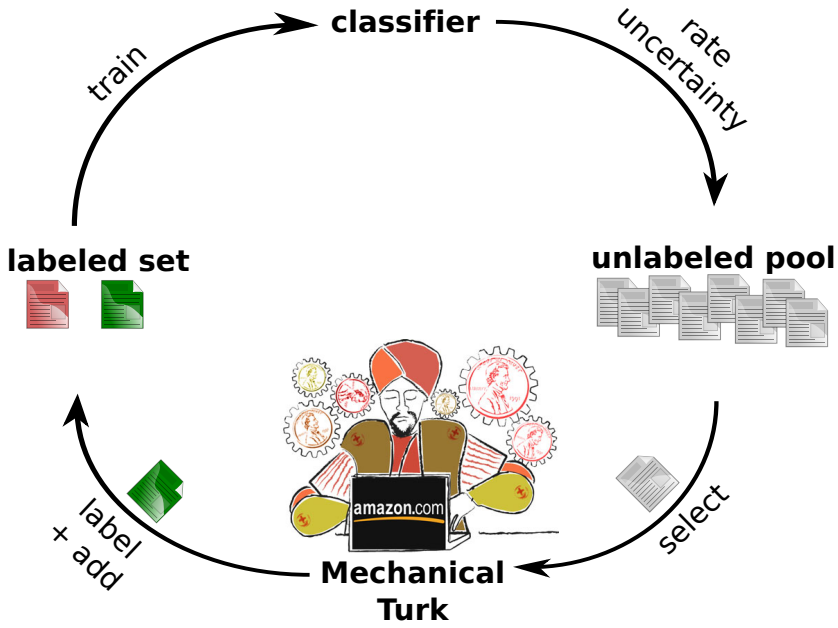
- Performance estimation for multiclass entity classification.
  - F-Score expectation.
  - Overoptimistic because of bad classifier probability estimates.
- Early stopping close to peak performance
- Two gradient-based stopping criteria:
  - Performance estimation gradient.
  - Uncertainty gradient.
- Reliable stopping at better-than-baseline performance.
- Saves annotation effort.

Outline

EMNLP 2011
joint work with Christian Scheible

Crowdsourcing

- Outsource annotation tasks to large crowd of workers.
- Mechanical Turk: popular platform.
- Low cost per annotation.
- Complementary to Active Learning.
- But: High noise.
- Will it work?



Standard Mechanical Turk workflow too static for AL.
Implement own system for managing and displaying tasks.

Annotation System – Features

- **Frontend**
  for posting and reviewing AL tasks on Mechanical Turk.
- **Concurrent example selection**: no wait time for annotators.
- Quality control: **Adaptive Voting**
  - Collect repeat annotations until agreement α.
  - If annotations disagree after d tries, discard example.
- **Fragment recovery** for NER:
  - Use agreeing parts of otherwise discarded sentence.
- **Play-by-play log** of user interactions.
  - for experiments with different parameters.

Experiments

NER: CoNLL-2003 Named Entity dataset

Sentiment: Movie Review sentiment classification

	NER	Sentiment
AL improves performance vs. Random	+6.9% F	+ 2.6 % Acc.
Voting improves performance vs. single labels	+ 3.5% F	+ 7.8 % Acc.
Voting improves label quality vs. single labels	+ 20 %	+ 13%

Robustness of Active Learning

- Noisy labels have big impact on trained classifier.
- But only small influence on selection: AL is robust!

Experiments

NER: CoNLL-2003 Named Entity dataset

Sentiment: Movie Review sentiment classification

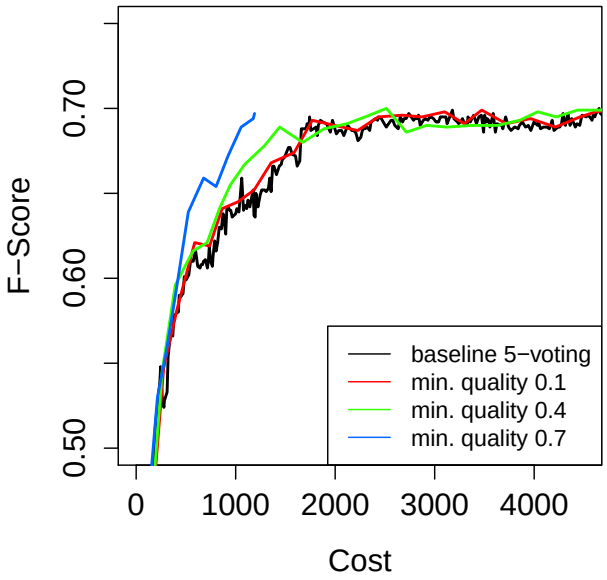
	NER	Sentiment
AL improves performance vs. Random	+6.9% F	+ 2.6 % Acc.
Voting improves performance vs. single labels	+ 3.5% F	+ 7.8 % Acc.
Voting improves label quality vs. single labels	+ 20 %	+ 13%

Robustness of Active Learning

- Noisy labels have big impact on trained classifier.
- But only small influence on selection: AL is robust!

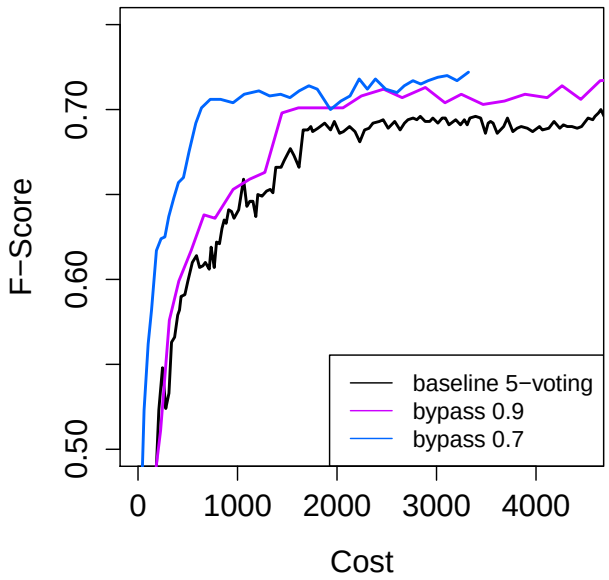
Using quality ratings for workers

Blocking low quality workers



- Low cutoffs help against spammers.
- High cutoffs achieve cost reduction.

Bypass voting for high quality workers

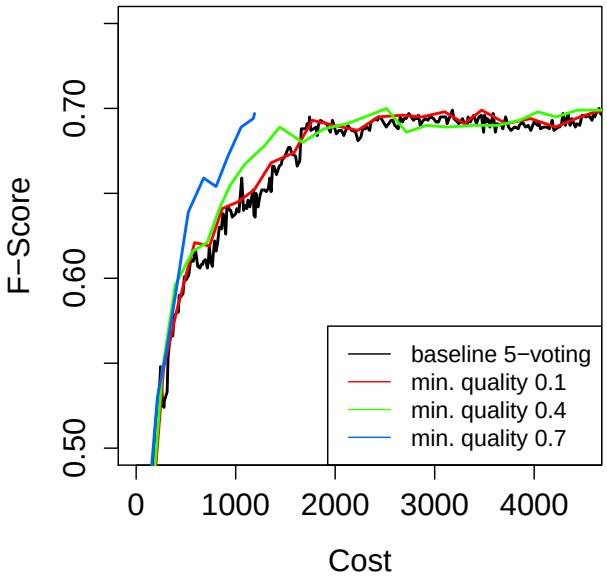


- Substantial cost reductions.
- Higher performance.

Oracle experiments: Quality ratings obtained using gold data.

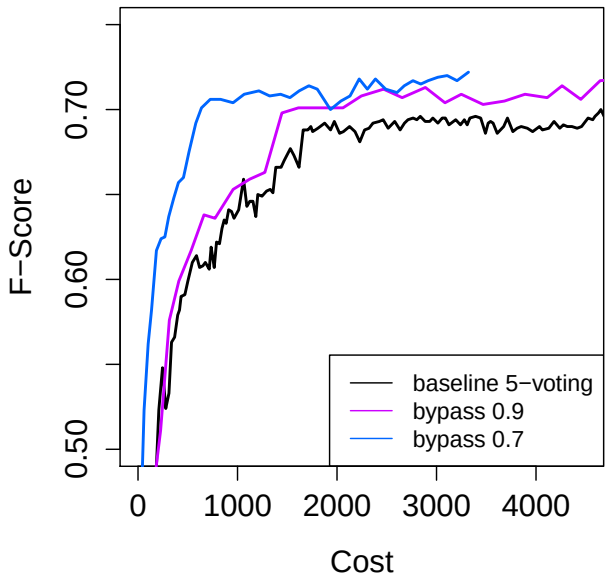
Using quality ratings for workers

Blocking low quality workers



- Low cutoffs help against spammers.
- High cutoffs achieve cost reduction.

Bypass voting for high quality workers



- Substantial cost reductions.
- Higher performance.

Oracle experiments: Quality ratings obtained using gold data.

Crowdsourcing - Summary

- Crowdsourcing offers low-cost annotation.
- But high noise level of annotations.

- Software to perform AL example selection on Mechanical Turk.
  - Concurrent retraining and example selection.
  - Adaptive Voting to control noise.
  - Fragment recovery.

- Active Learning is effective with Crowdsourcing.
- Voting improves label quality.
- Active Learning can be robust to noise.

- Quality ratings can further improve performance.

Summary

- Active Learning: Interactive method to reduce annotation effort.
- Example selection strategies
  for NER and Coreference resolution.
  - NER: Sentence selection effectively addresses
    missed class problem.
  - Coreference: First successful application of Active Learning.
- Stopping criteria successfully identify peak performance.
- First successful application of Active Learning to
  Crowdsourcing.