A Cross-corpus Study of Subjectivity Identification Using Unsupervised Learning

Yang Liu
The University of Texas at Dallas

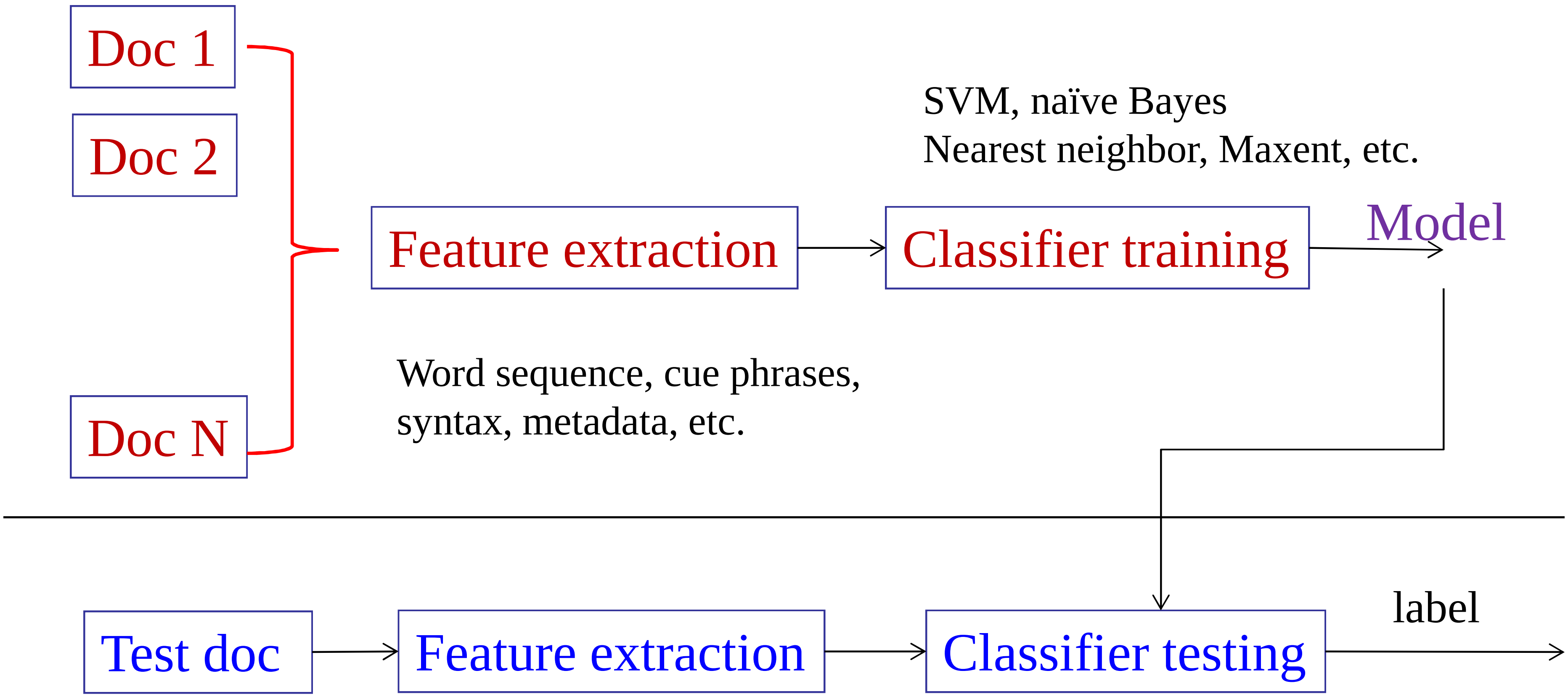
Sentire 2013

Acknowledgment: Dong Wang

Introduction

- Increasing interest in sentiment analysis
  - Data: reviews, news article, blogs, tweet, youtube …
  - Approach: various models, different levels of information
  - Classification level: word, document, aspect/features
  - Task definition: polarity, subjectivity, emotion, speaker/writer vs. listener/reader
  - End task goal: business intelligence, stock, poll…

Supervised learning



Problems with supervised classifiers

- Supervised learning requires annotated training data
  - Lack of data for many domains

- Mismatched training and test conditions
  - Differences in domain/genre, style, class labels, etc.

Example of a new domain: speech

- Large amount of speech data that contains sentiment/affect
  - Talk shows, debates, conversations, meetings, etc.
- Speech contains rich information about speakers’ affective states





Speech example

D : could the middle button of the on-screen menu function as a power button? [pos-sub]

C: um not really, [neg-sub]

C: it would make it hard to turn the machine off, to turn your TV off. [neg-sub]

A: mm-hmm [obj]

B: if you pressed and held it maybe. [pos-sub]

C: yeah, yeah, that that'd be one way of doing it, yeah. That'd work, yeah. [pos-sub]

D: if you like held it down, that would be on off. [pos-sub]

B: yeah. On off, that's a possibility, yeah. [pos-sub]

A: okay. [obj]

This talk

- Goal of this study: subjectivity detection across different domains
- What is the domain difference?
- Can unsupervised or semi-supervised learning help?
- What are the impacting factors?

Data: AMI

- AMI meeting
  - Multiparty meeting corpus (role playing scenario)
  - Classification units based on dialogue act labels (DA).

  - Example

    - It does make sense from maybe the design point of view. (SUBJECTIVE)
    - My task was this time to put up a questionnaire. (OBJECTIVE)

Data: movie data

- Movie data (Pang and Lee 2004)
  - Subjective sentences from movie reviews and objective sentences from movie plot summaries
  - Example
    - It's hard to tell with all the crashing and banging where the salesmanship ends and the movie begins. (SUBJECTIVE)
    - The movie begins in the past where a young boy named Sam attempts to save celebi from a hunter. (OBJECTIVE)

Data: MPQA

- MPQA corpus (Wilson and Wiebe 2003)
  - Sentences from news articles and labeled by human.

  - Example

    - The world community should not tolerate crime of war. (SUBJECTIVE)
    - The European Commission announced it had pledged a nancial package of grants and loans totaling 530 million euros (450 million dollars). (OBJECTIVE)

Data statistics

		Movie	MPQA	AMI
# of sents	subjective	5,000	5,000	4,946
	objective	5,000	5,000	4,946
sent length	min	3	1	3
	max	100	246	67
	mean	20.37	22.38	8.78
	variance	75.26	147.18	34.26
vocabulary size		15,847	13,414	3,337
Inter-annotator agreement		**N/A**	**0.77**	**0.56**

Unsupervised learning approach

- Create initial training set: use a subjective lexicon to calculate subjectivity score for each sentence/DA.

$$sub(s) = (\sum_{w \in s} sub(w)) / length$$

$$sub(s) = (\sum_{w \in s} sub(w)) / \log(length)$$

- Evaluate two semi-supervised methods to iteratively learn from unlabeled data
  – Self-training
  – Calibrated EM

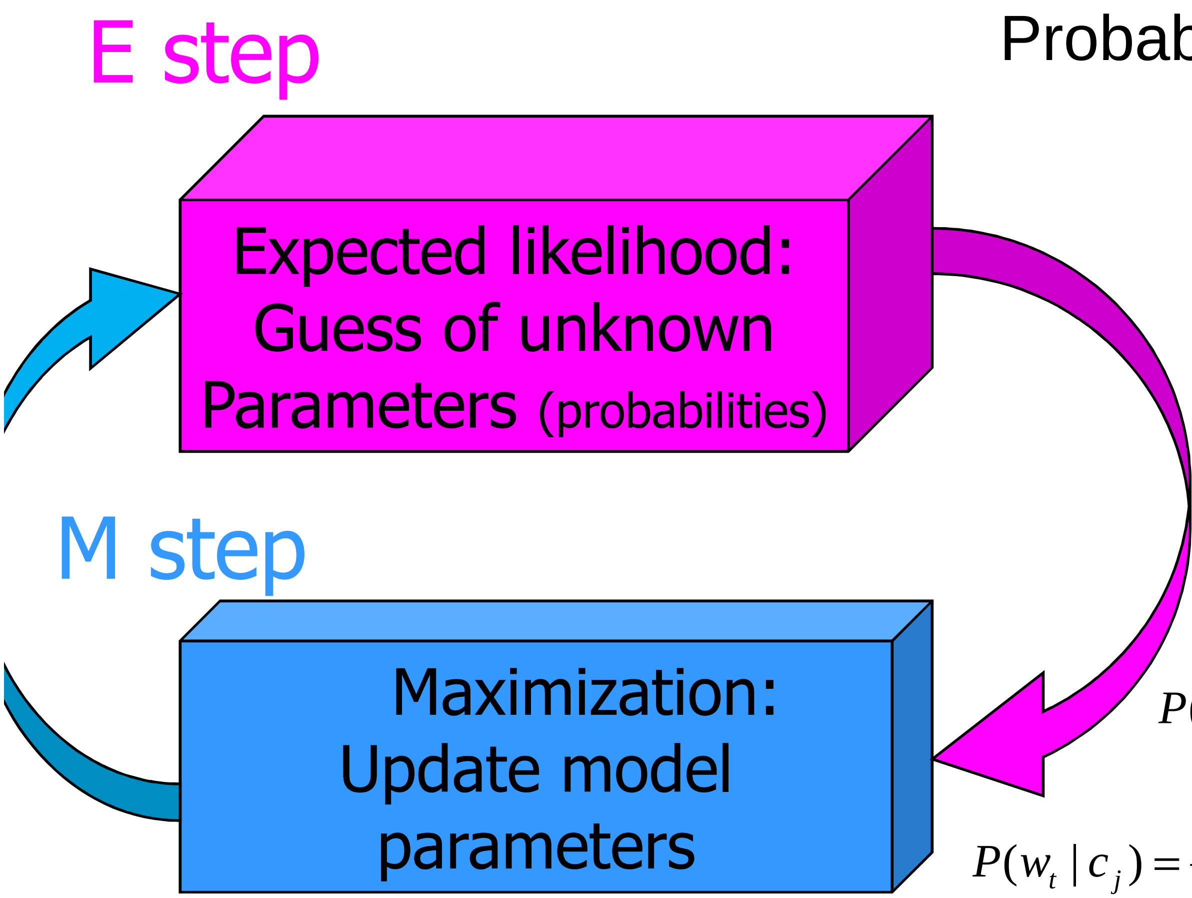
Self-training

- Assumption: one’s own high confidence prediction is correct
- Algorithm:
  - Train classier using initial labeled data
  - Use trained classier to label unlabeled data
  - Add top ranked n subjective and n objective examples to training data, remove from unlabeled
  - Repeat

Self-training

- Advantage
  - Simple method
  - Applies to any classifiers
- Disadvantage
  - Early mistakes may have a negative impact, can't remove added labeled examples
  - No guarantee on convergence

Basic EM for Naïve Bayes classifier



Probability of sentence in class

$$P(c_j \mid s_i) = \frac{P(c_j)P(s_i \mid c_j)}{P(s_i)}$$

$$= \frac{P(c_j)\prod_{k=1}^{|s_i|} P(w_k \mid c_j)}{\sum_{c_l \in C} P(c_l)\prod_{k=1}^{|s_i|} P(w_k \mid c_l)}$$

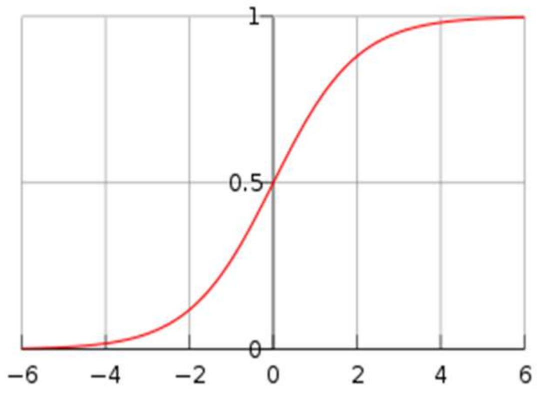
NB probabilities

$$P(c_j) = \frac{0.1 + \sum_{s_i \in S} P(c_j \mid s_i)}{0.1 \times |C| + |S|}$$

$$P(w_t \mid c_j) = \frac{0.1 + \sum_{s_i \in S} N(w_t, s_i) P(c_j \mid s_i)}{0.1 \times |V| + \sum_{k=1}^{|V|} \sum_{s_i \in S} N(w_k, s_i) P(c_j \mid s_i)}$$

Calibrated EM

- Problem with Naïve bayes: posteriors are not accurate, tend to be close to 0 or 1
- Calibrated EM (Tsuruoka and Tsujii 2003):
  – shift posterior probability *p* of unlabeled data to generate desired class distribution.
    - p' = inverse_sigmoid(p)
    - p'= p' – median of p'
    - p = sigmoid (p')



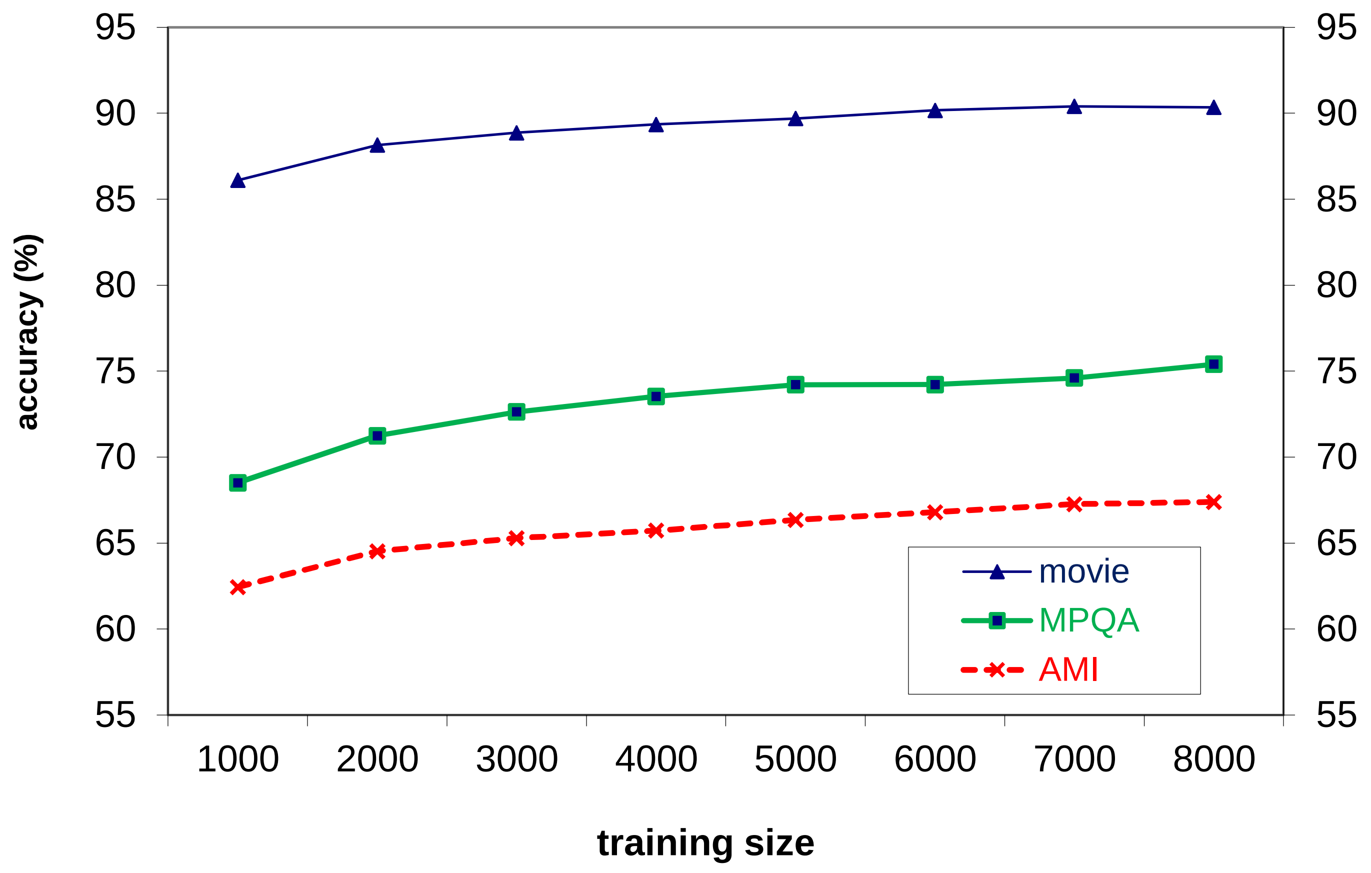
EM for naïve Bayes

- Advantage
  - Clear probabilistic framework
  - Can be effective if the model is close to correct
  - No hard decisions for added samples
- Disadvantage
  - Model may not be correct
  - Local optima in EM
  - Added samples may hurt performance

Experimental setup

- Use unigrams as features (bag-of-words model)
- 5-fold cross validation
  - divide the corpus into 5 parts
  - in each run, reserve one part as test set, and treat the rest as unlabeled data.

Supervised results

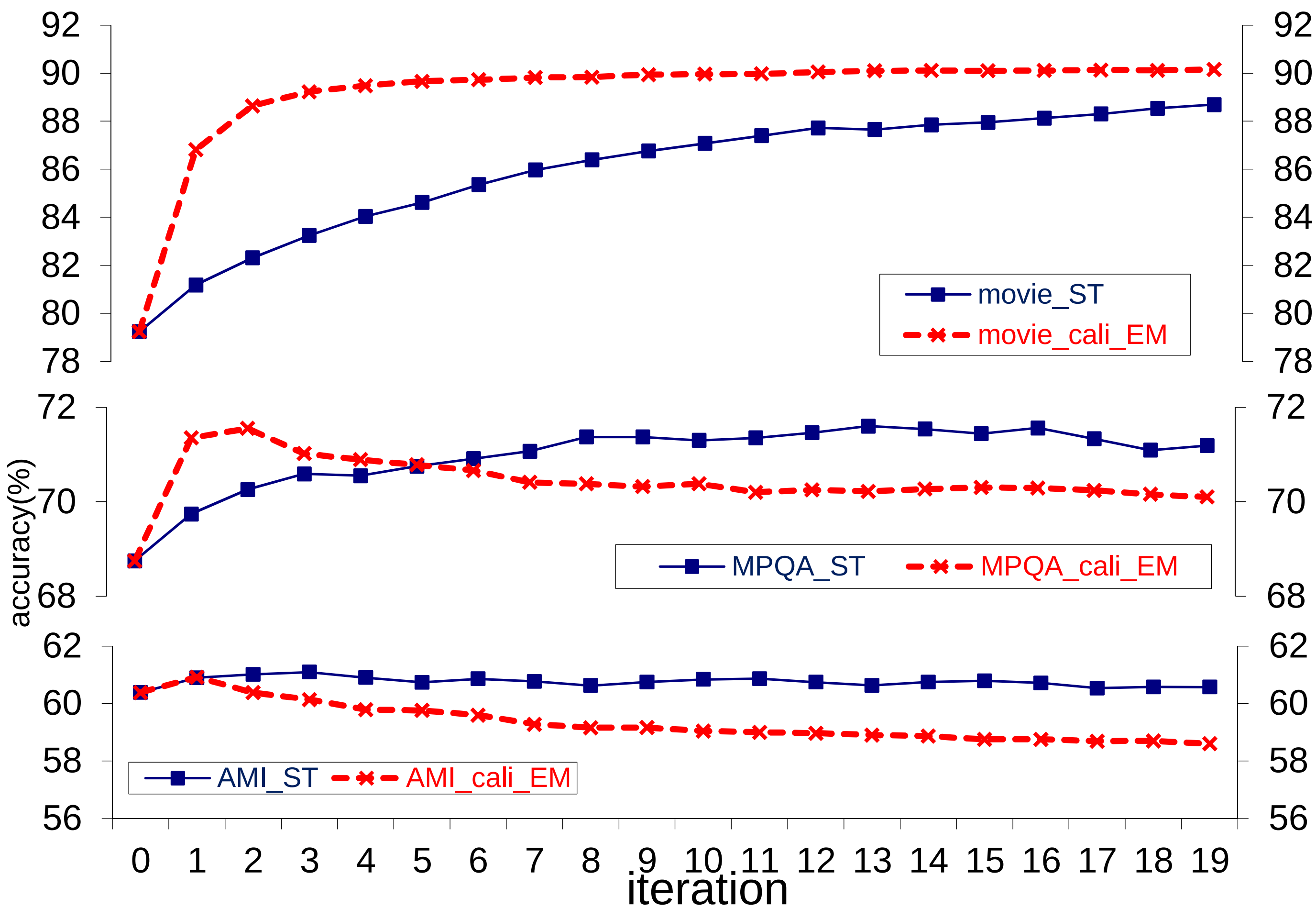


Unsupervised results

2k initial instances

Movie and MPQA

AMI



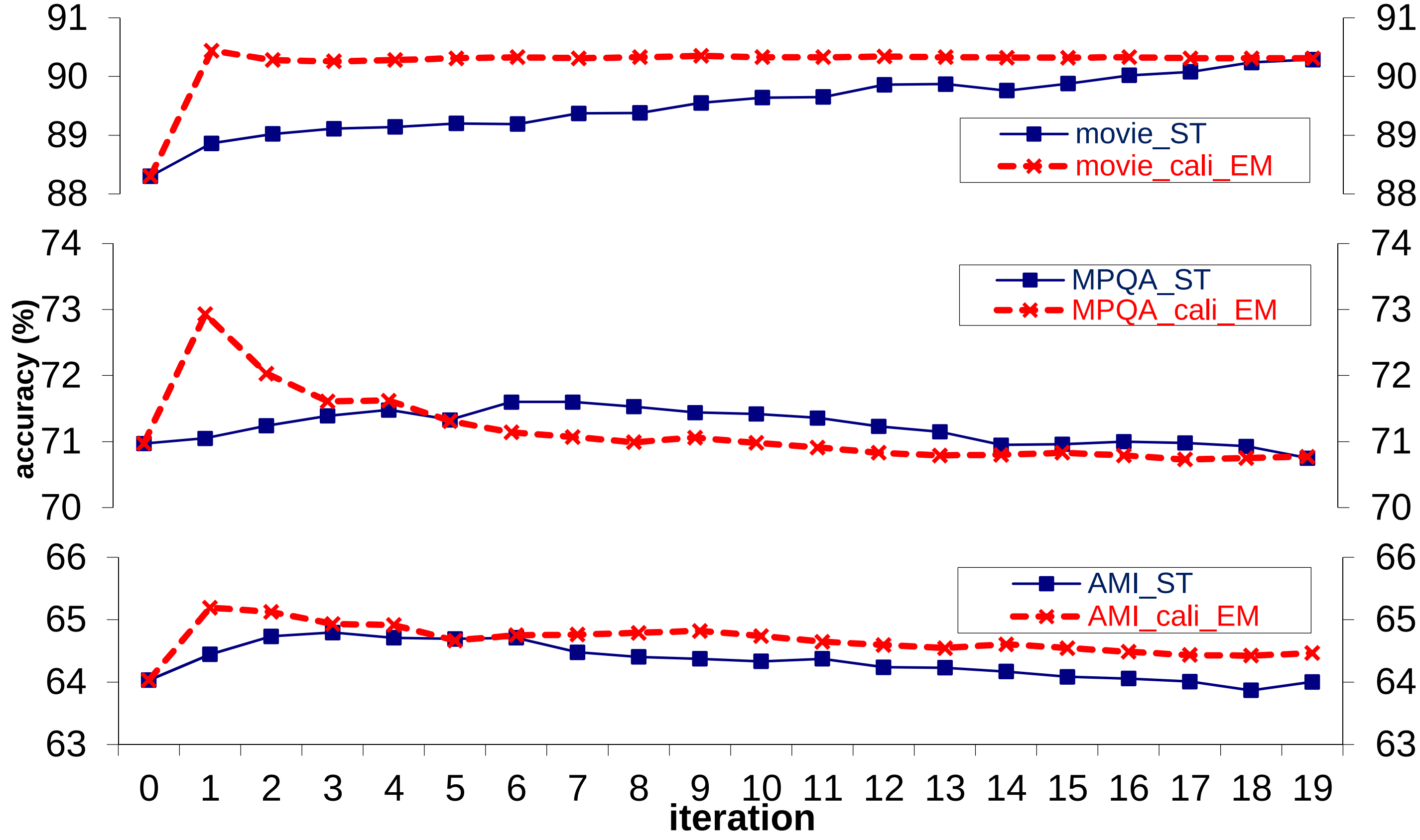
Analysis: initial training set

- How does the accuracy and size of the initial training set affect performance?

size	movie			MPQA			AMI		
	sub	obj	Acc On test	sub	obj	Acc On test	sub	obj	Acc On test
100	95.20	82.20	59.93	83.20	87.60	60.45	49.60	71.60	50.51
200	90.20	82.00	71.63	85.60	86.60	63.83	53.40	71.00	53.81
1000	82.48	80.88	77.62	85.76	87.64	66.98	65.96	68.56	60.53
2000	79.24	79.04	79.24	85.18	87.46	68.75	66.98	69.04	60.39
3000	77.13	77.31	79.64	82.53	85.92	70.05	67.05	69.89	60.46

Analysis: semi-supervised setting

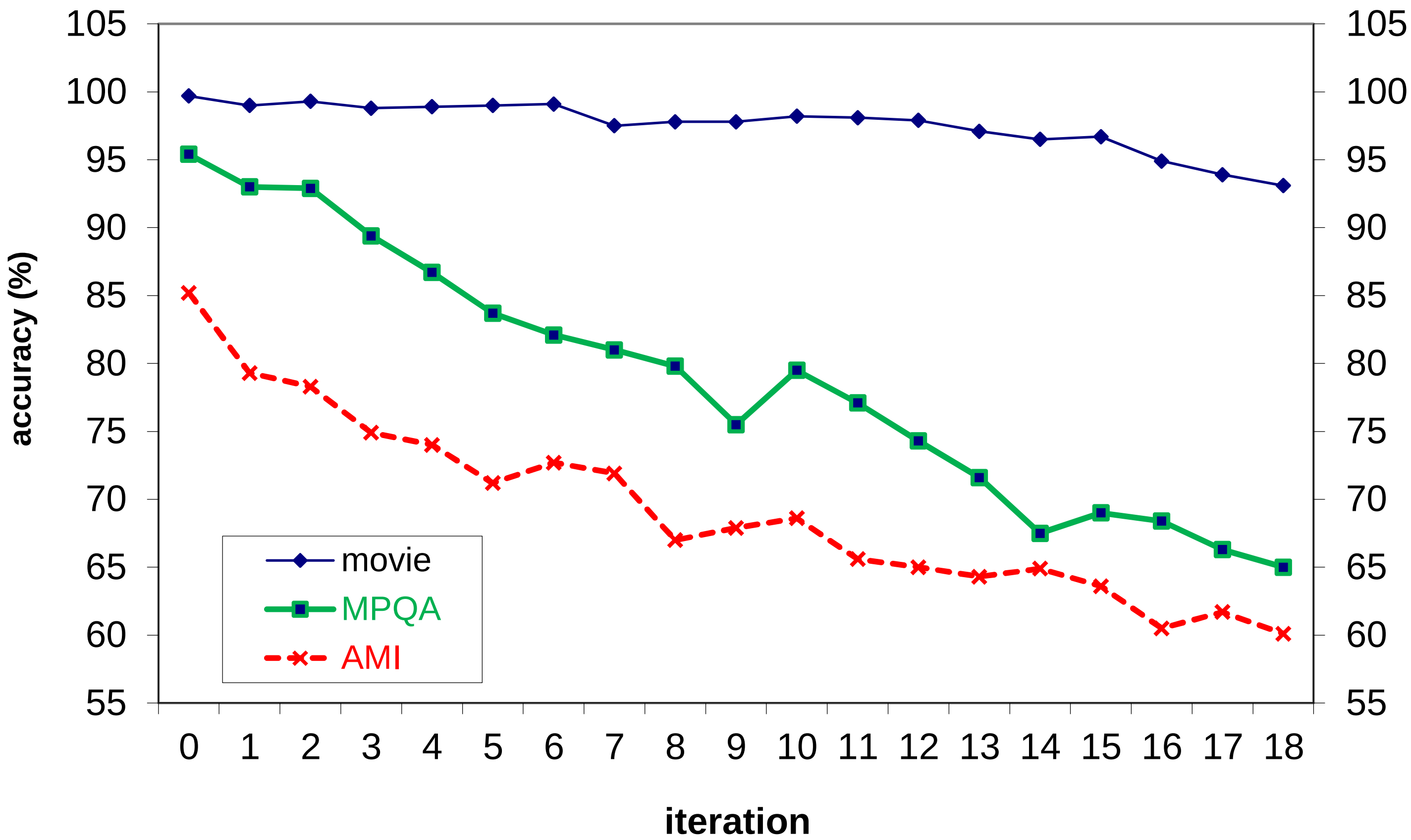
- What if we use labeled data as initial training set, i.e., semi-supervised learning?



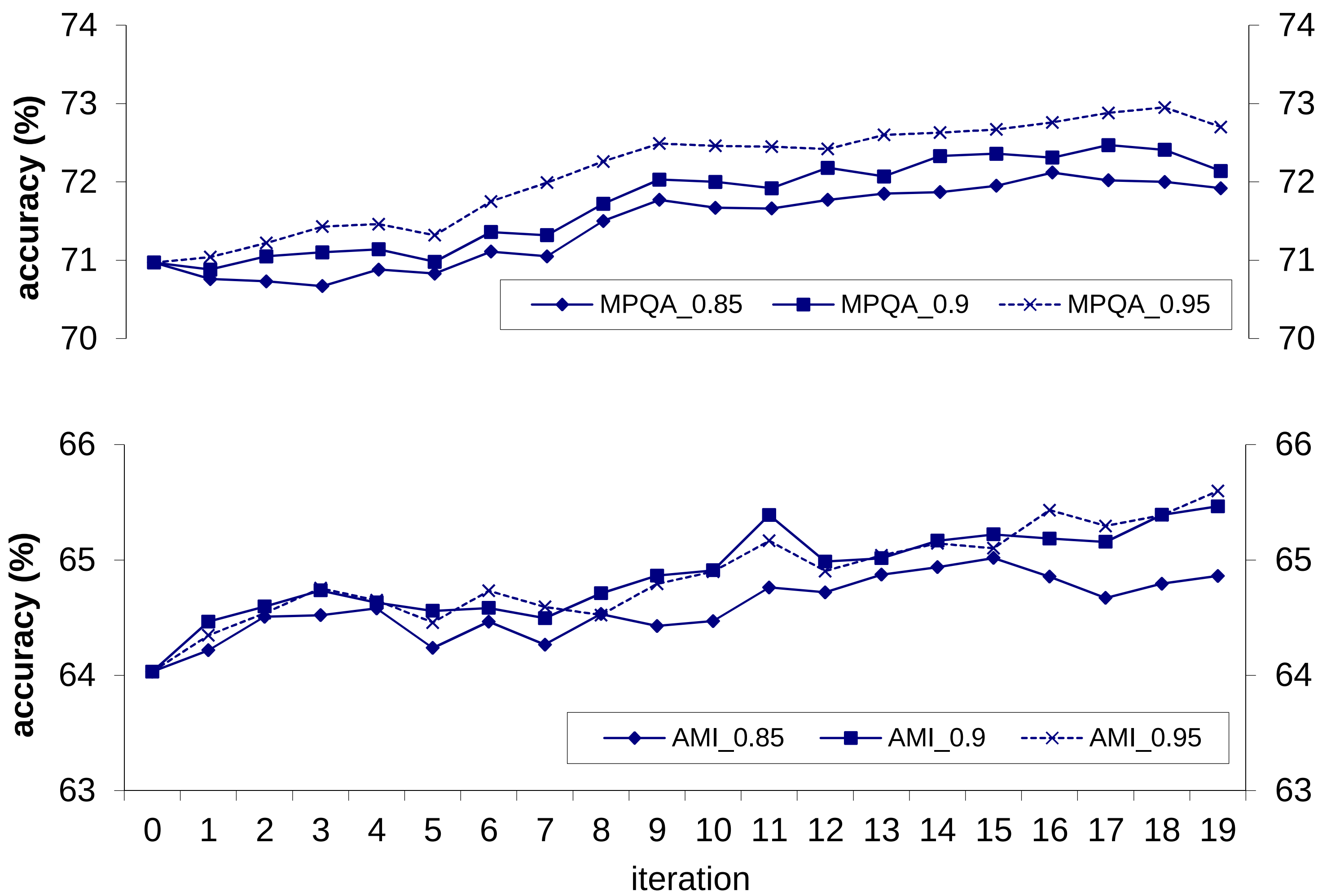
2k labeled instances

Self-training analysis: accuracy of added examples

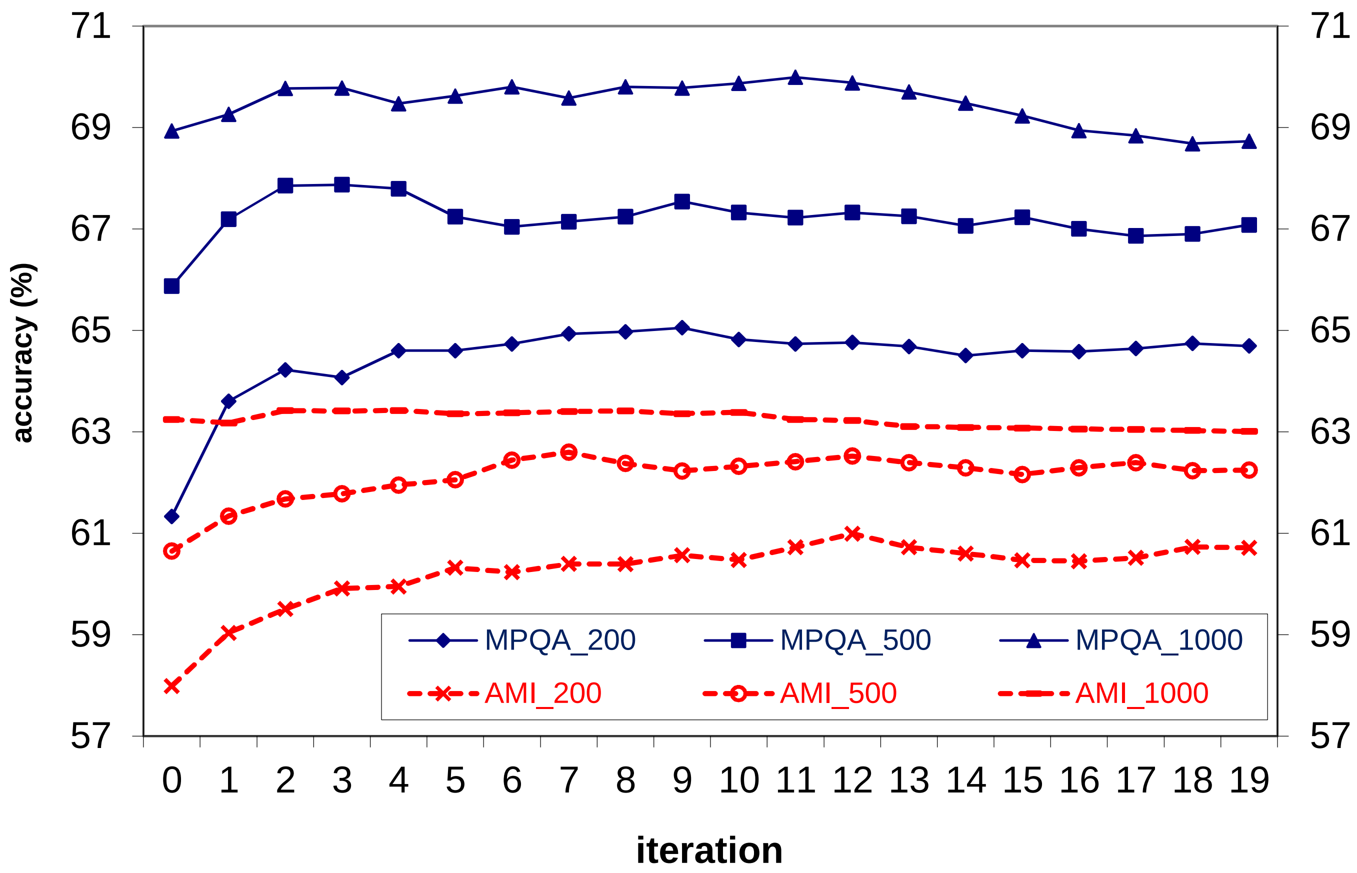
Semi-supervised setup



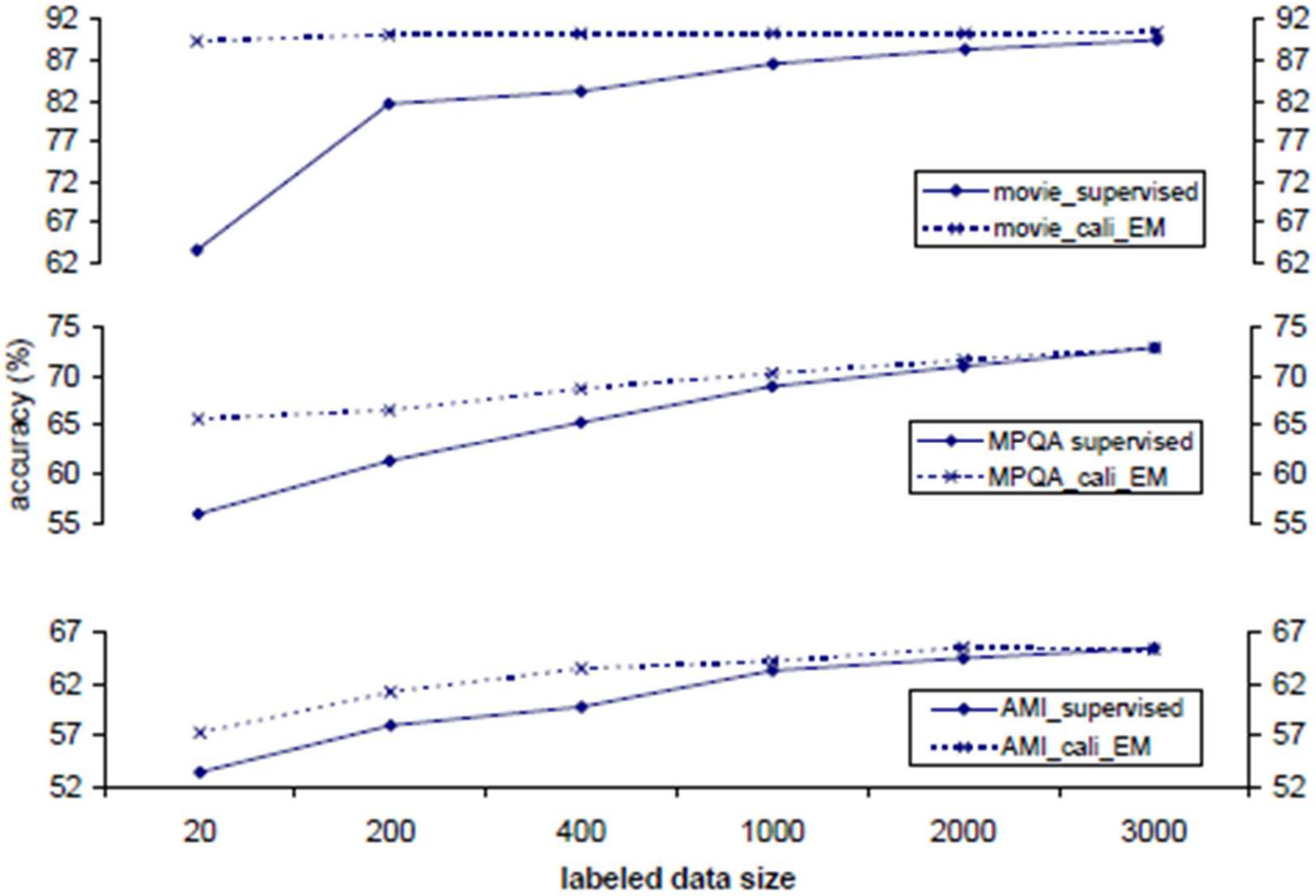
Self-training analysis: control added example accuracy



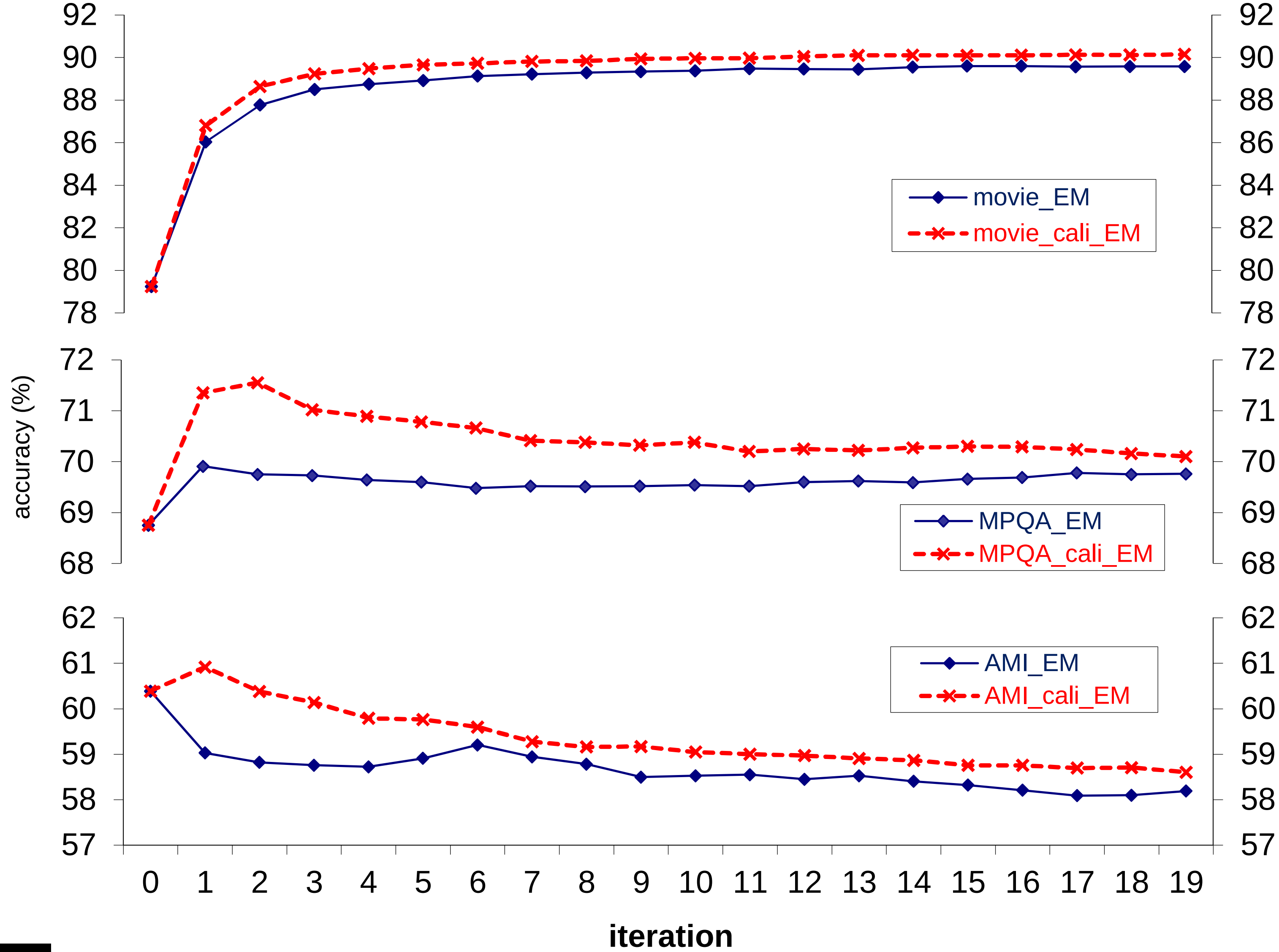
Self-training analysis: different size of initial data



EM analysis: different initial size



EM analysis: effect of calibration



Summary of results

- Observe significantly different patterns in speech data vs. other two corpora.
- The base classifier performance has a substantial impact on iterative learning.
- For corpora with low classification accuracy, the bootstrapping methods are useful only when the initial training size is small and initial accuracy is low.

Discussions

- Class distribution
  - Similar observation on imbalanced data
  - However, assumed distribution is known
- Domain difference
  - Vocabulary, sentence length, error patterns on subjective and objective sentences
- Model limitations
  - Bag of words
  - Expect similar patterns when changing the baseline learning approach (?)

Improving sentiment analysis on speech data

- Sentiment analysis is hard on spoken text
- How can we improve its performance?
  - Increase annotated training data
  - Domain adaptation
  - Design domain specific models/features
    - Previous studied investigated using acoustic/prosodic cues in sentiment analysis

Other work

- Summarization of speaker's opinion
  - Used Switchboard conversations
- Emotion recognition from speech

- Automatic summarization
  - News article, meetings, social media
- Text normalization in social media
- Language processing in clinical applications


Thank you!