UNIVERSITÄT
BERN

Semi-Supervised Learning for Handwriting Recognition

Volkmar Frinken
University of Bern

frinken@iam.unibe.ch

31. August 2009

IM2 Summer Institute, Chavannes-de-Bogis

UNIVERSITÄT BERN

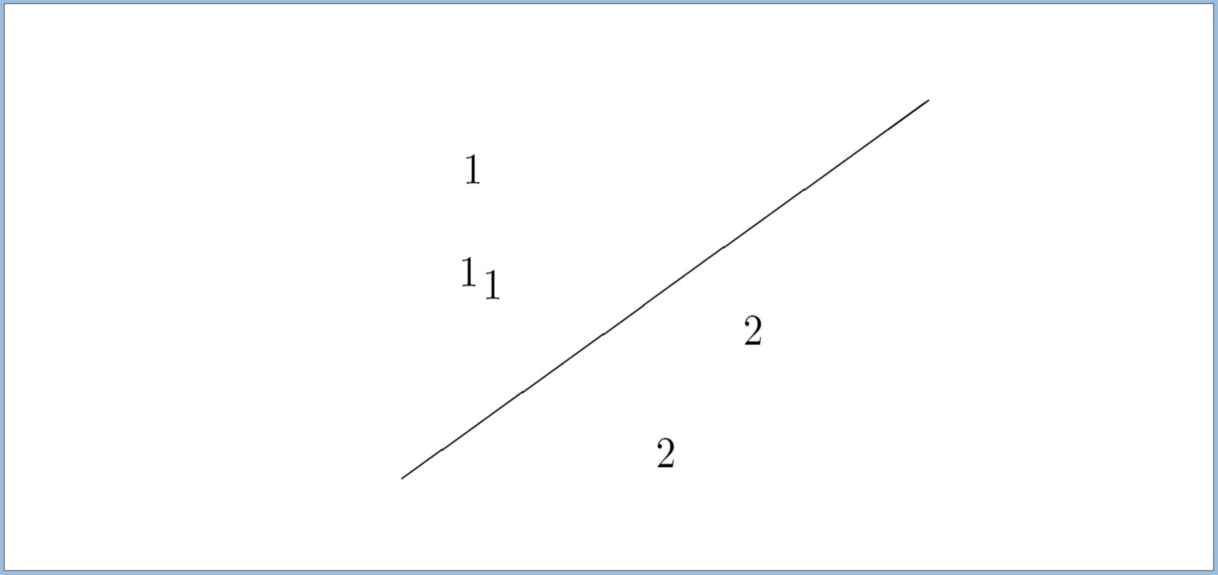
Semi-Supervised Learning

> Unsupervised Learning (Clustering)
  — Aim: To identify structures in the feature space
  — Prerequisites: Set of training elements

> Supervised Learning (Classification)
  — Aim: To find a mapping from feature space to label space
  — Prerequisites: Set of labeled training elements

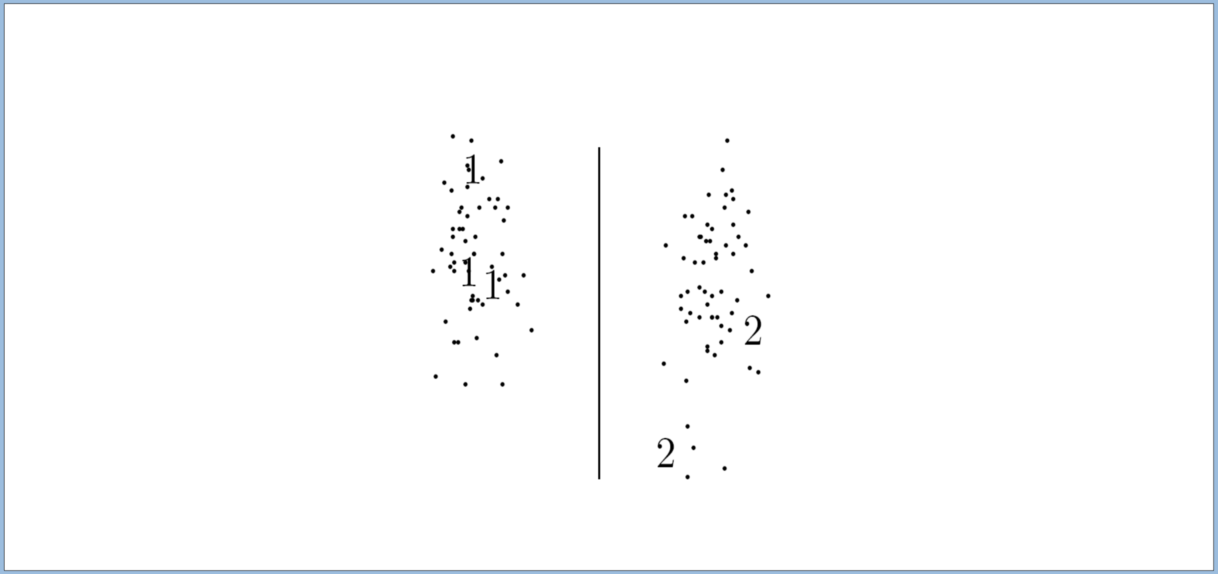
Semi-Supervised Learning

- Semi-Supervised Learning (Classification)
  - Exploit information about the feature space from unlabeled data to find a mapping from feature space to label space
  - Prerequisites: Set of labeled and unlabeled training elements



Semi-Supervised Learning

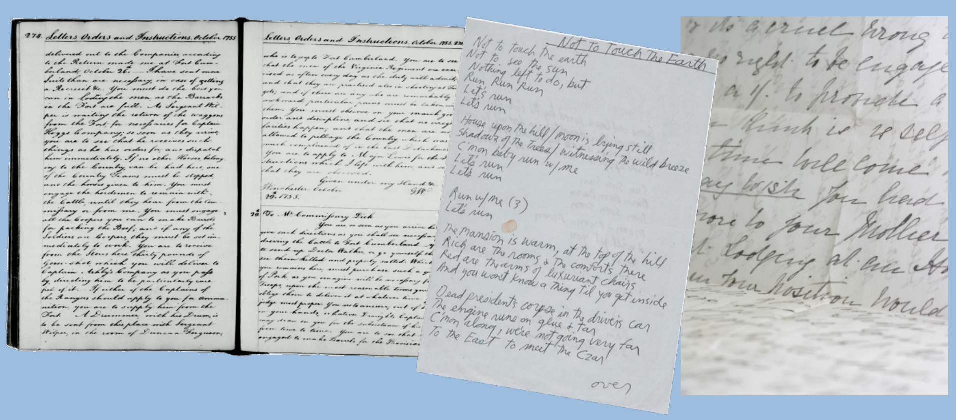
- Semi-Supervised Learning (Classification)
  - Exploit information about the feature space from unlabeled data to find a mapping from feature space to label space
  - Prerequisites: Set of labeled and unlabeled training elements



Why Semi-Supervised Learning

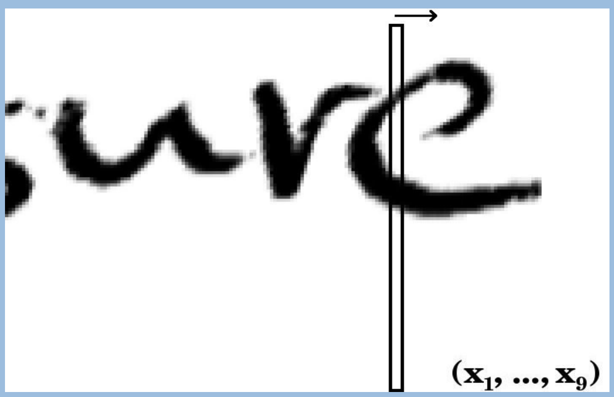
- Unlabeled Data might be cheap to acquire and vastly available while labeled data is rare and costly

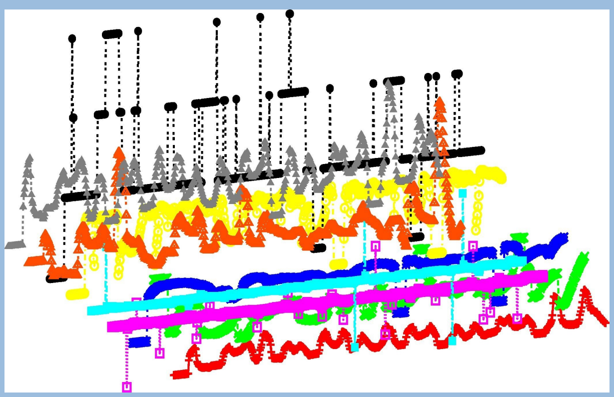
- In the case of Handwriting Recognition:
  - — the ground truth has to be created manually
  - — unlabeled handwritten text is nearly everywhere



Handwriting Recognition

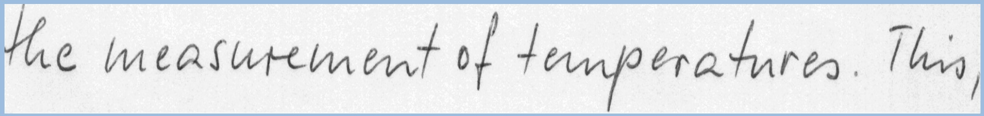
- Data
  - IAM Handwriting Data Base
- Features
  - Extracting Features using a sliding window approach
  - Result: Sequence of 9-dim vectors

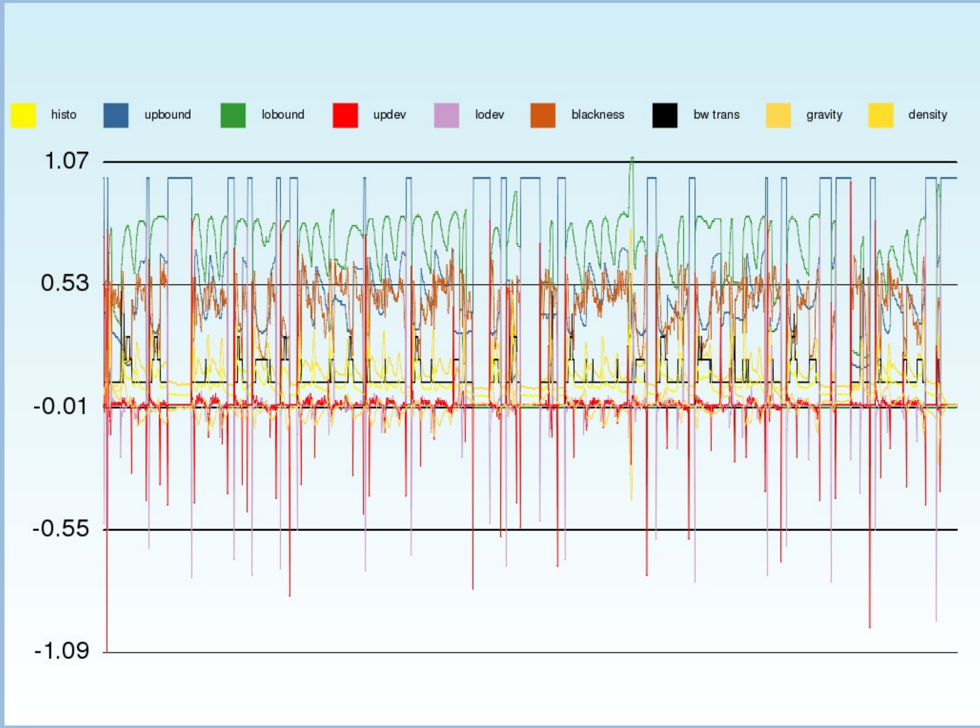




Sequential Data

- No vectorial description
- Lack of algorithmic tools
- No clear separation between different elements
- No easy distance measure
- Classification with an exponentially growing class label space





How Semi-Supervised Learning

UNIVERSITÄT BERN

- Self-Learning
  - A recognizer is trained on the labeled set
  - The recognizer decodes the unlabeled set
  - Confident recognitions are used to create a new training set

- Co-Learning
  - Two (different) recognizers are trained on the labeled set
  - The recognizer decodes the unlabeled set
  - Confident recognitions are used to create a new training set for the other recognizer

Theoretical Background Self-Learning

- Self-Learning is a form of Expectation-Maximization
- EM maximizes the likelihood of incomplete data fitting to a model
- EM iterates two steps
  - E-step: Computing the expectation of the incomplete data according to the model
  - M-step: Selecting new model parameters

Theoretical Background Self-Learning

b UNIVERSITÄT BERN

- Missing labels can be seen as incomplete data

- Self-Learning iterates two steps
  - E-Step: Computing the missing labels
  - M-Step: Retrain the recognizer using these new labels

Theoretical Background Self-Learning

b
UNIVERSITÄT
BERN

EM requires the expectation value of incomplete data

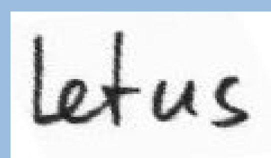
But:
computing the expectation value of labels is not possible

Theoretical Background Self-Learning

EM requires the expectation value of incomplete data

But:
computing the expectation value of labels is not possible

Example:



“Let us”	80% probability
“Lotus”	17% probability
“Lexus”	03% probability

0.8 “Let us” + 0.17 “Lotus” + 0.03 “Lexus” = ???

Theoretical Background Self-Learning

We select only those words, that are recognized with a high confidence and use that label

“Let us”	80% probability	→	“Let us”	100% probability
“Lotus”	17% probability		“Lotus”	0% probability
“Lexus”	03% probability		“Lexus”	0% probability

“Let us”	60% probability	→	Ignore word
“Lotus”	30% probability		
“Lexus”	10% probability		

b UNIVERSITÄT BERN

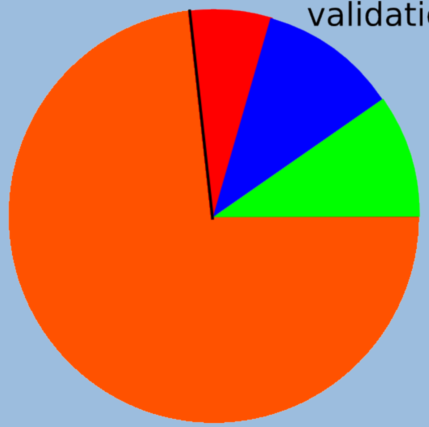
Self-Learning

What is the best threshold?

> High Threshold
  — Nearly no errors in the labels
  — Very few data elements enter the new training set
  — Retraining set nearly does not change
  —
> Low Threshold
  — More data elements enter the training set
  — Uncertain recognitions include error
  — Erroneous labels in the training set impede the performance

Experimental Evaluation

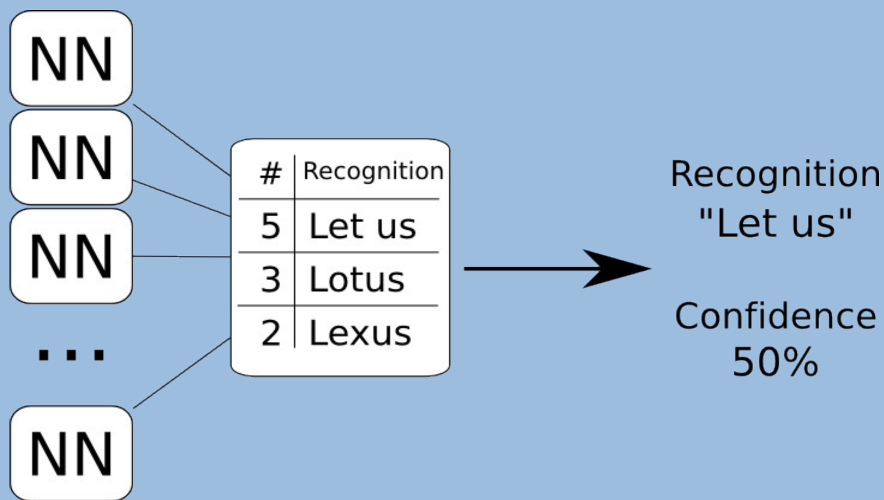
- Single word recognition
- Recognizer Recurrent Neural Networks
- Data set
  - 4'000 most frequent words from the IAM data base
  - 4 sets
    - Test set: 5,342 words from 52 writers
    - Validation set: 5,590 word from 56 writers
    - Work set: 38,127 words from 238 writers
      - Split up into labeled train set and unlabeled set



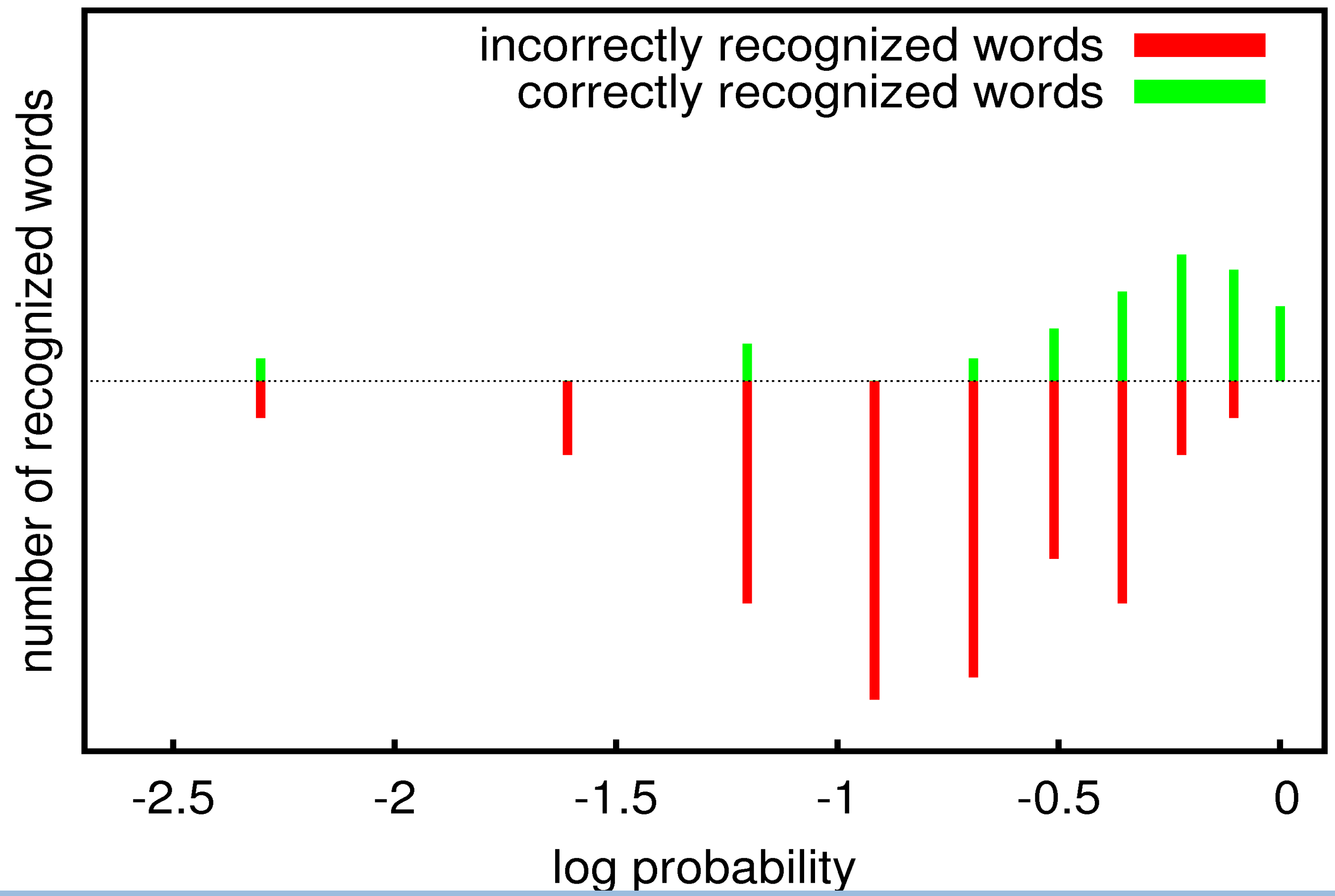
- Two recognition modes
  - Recognition with and without a dictionary

Recognition Confidence

- The Confidence returned by the Neural Networks is not reliable
- Train of several Neural Networks
- Majority voting on the outputs
- Number of agreeing Networks are used as recognition confidence



Confidence based Rules for Selecting Elements for Retraining



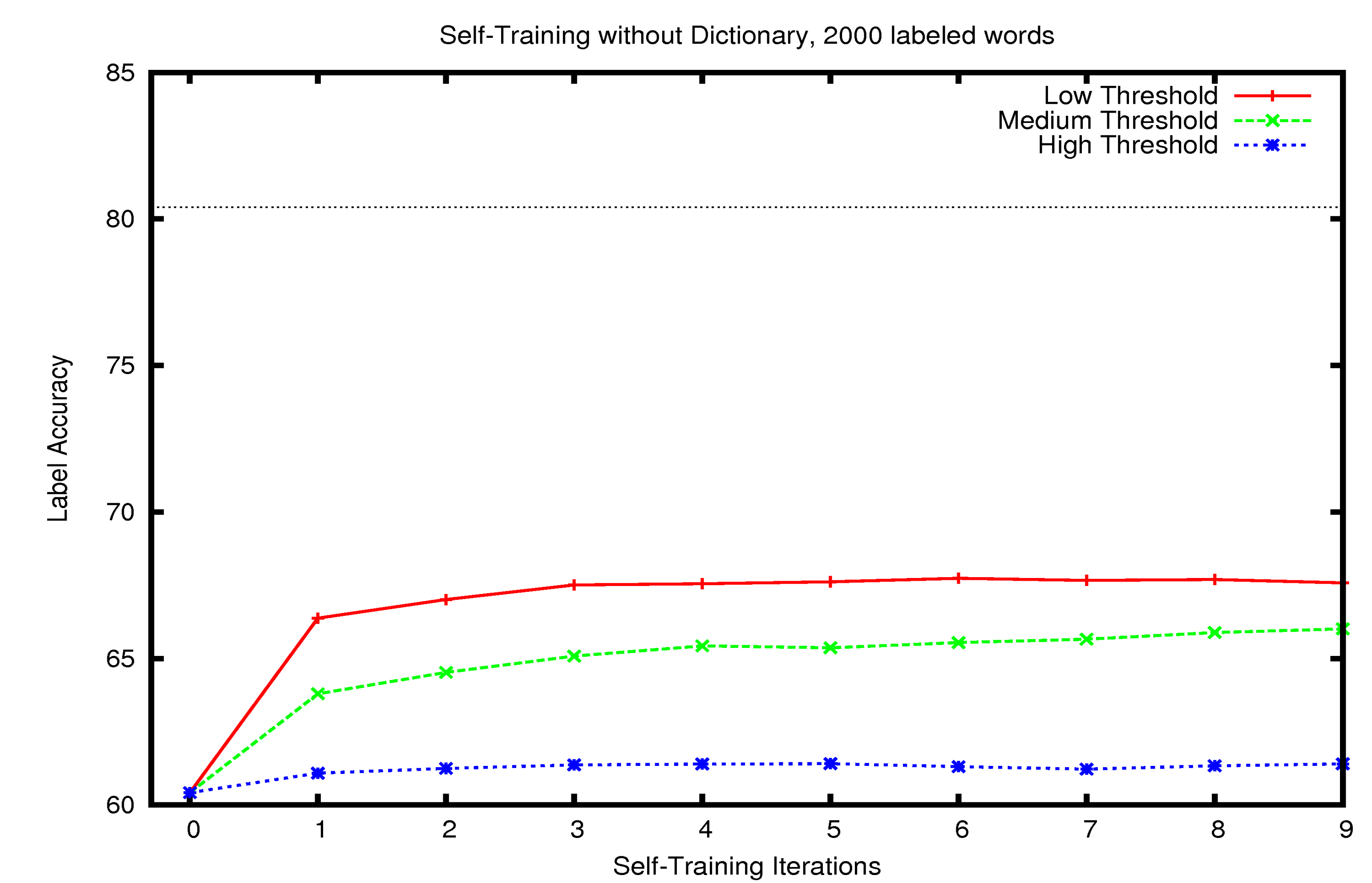
UNIVERSITÄT BERN

Confidence based Rules for Selecting Elements for Retraining

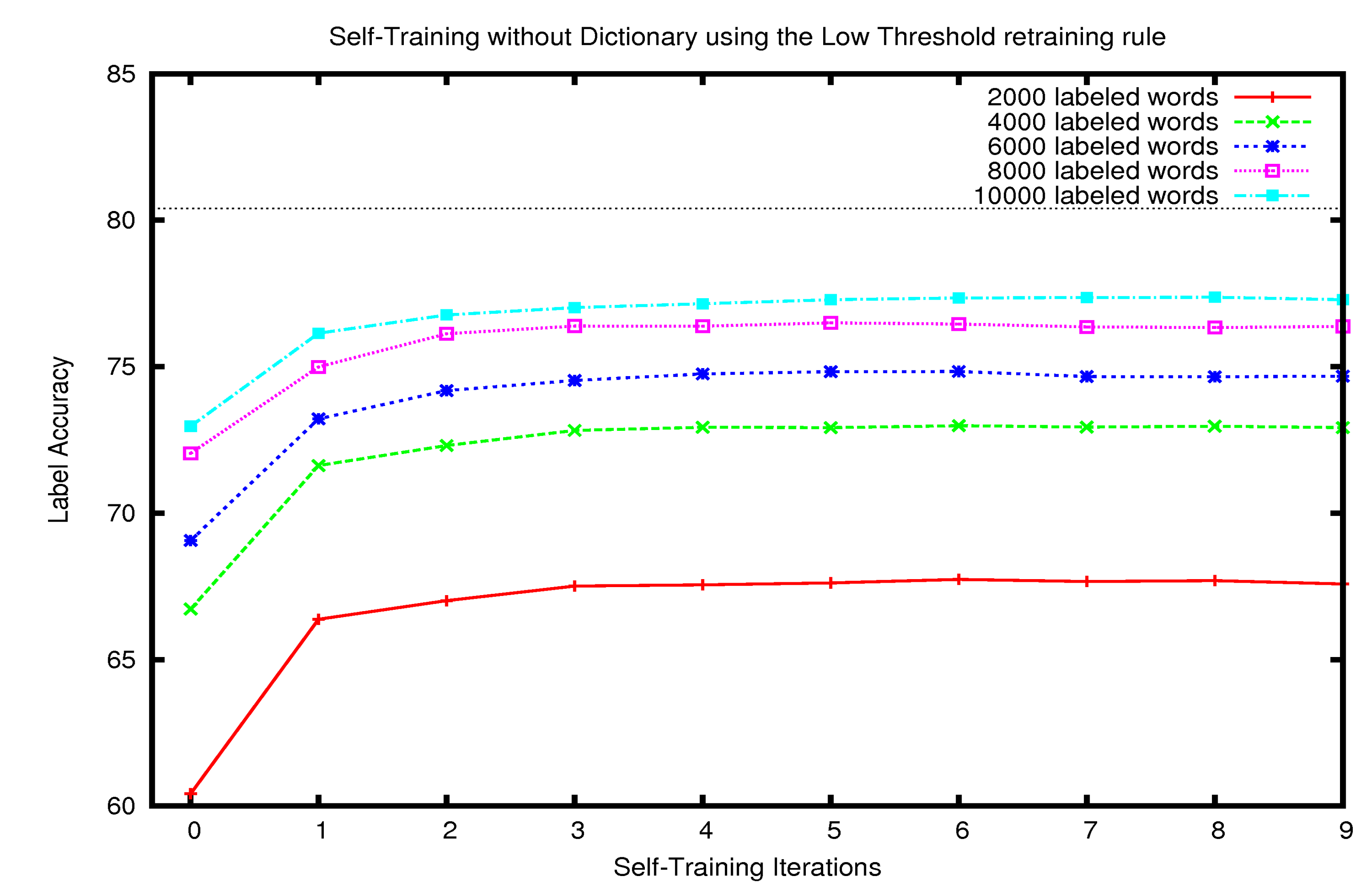
- We investigated different retraining rules
- The precise thresholds were set in each iteration according to the validation set
- All words having a confidence higher that the threshold were selected for retraining

- High Threshold Retraining Rule
  - Threshold set to lowest value without errors
- Medium Threshold Retraining Rule
  - Threshold set to lowest values with more correct than incorrect recognitions
- Low Threshold Retraining Rule
  - Threshold set to -infinity

2000 labeled words, no dictionary



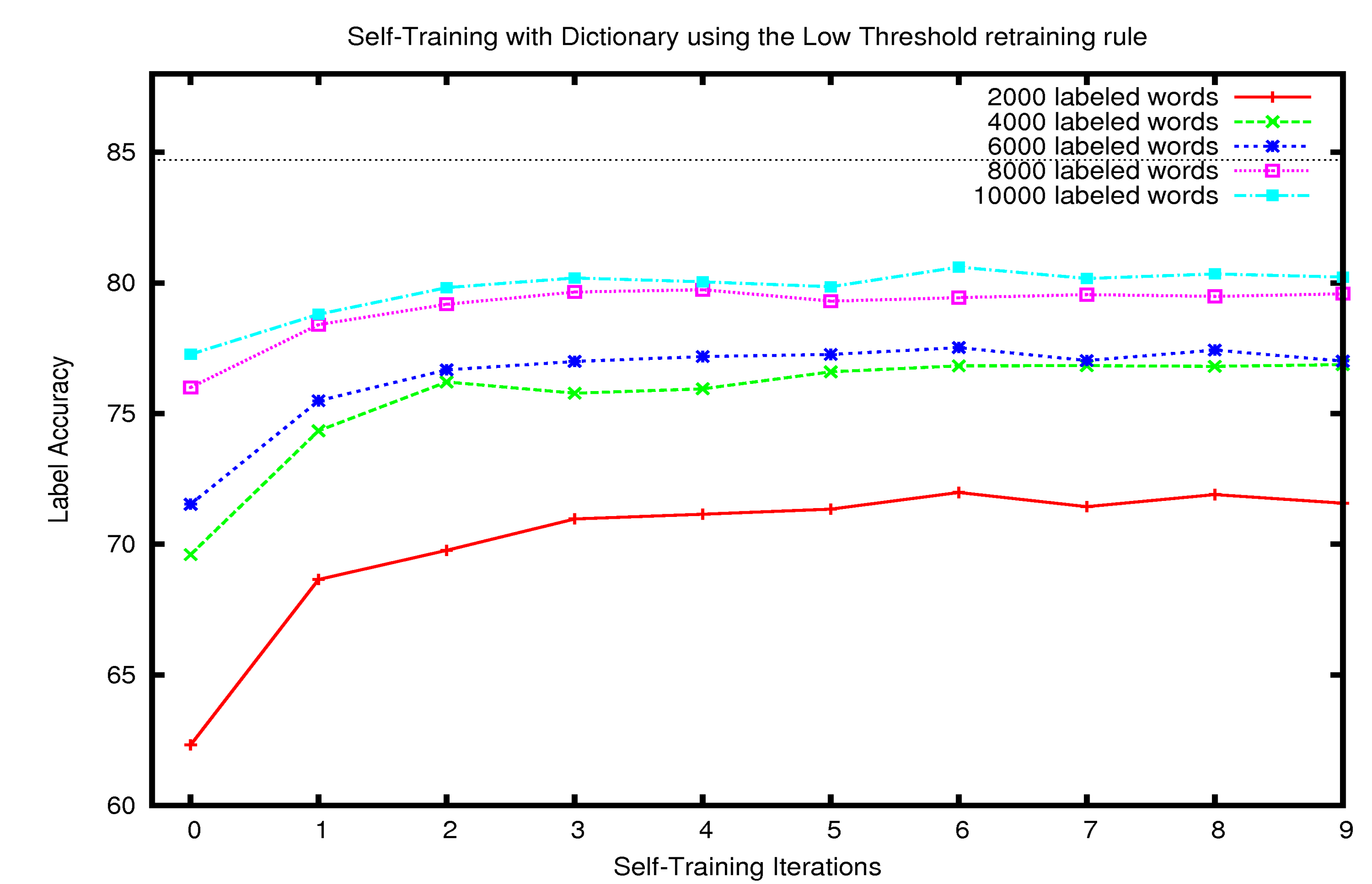
Low Threshold Retr. Rule



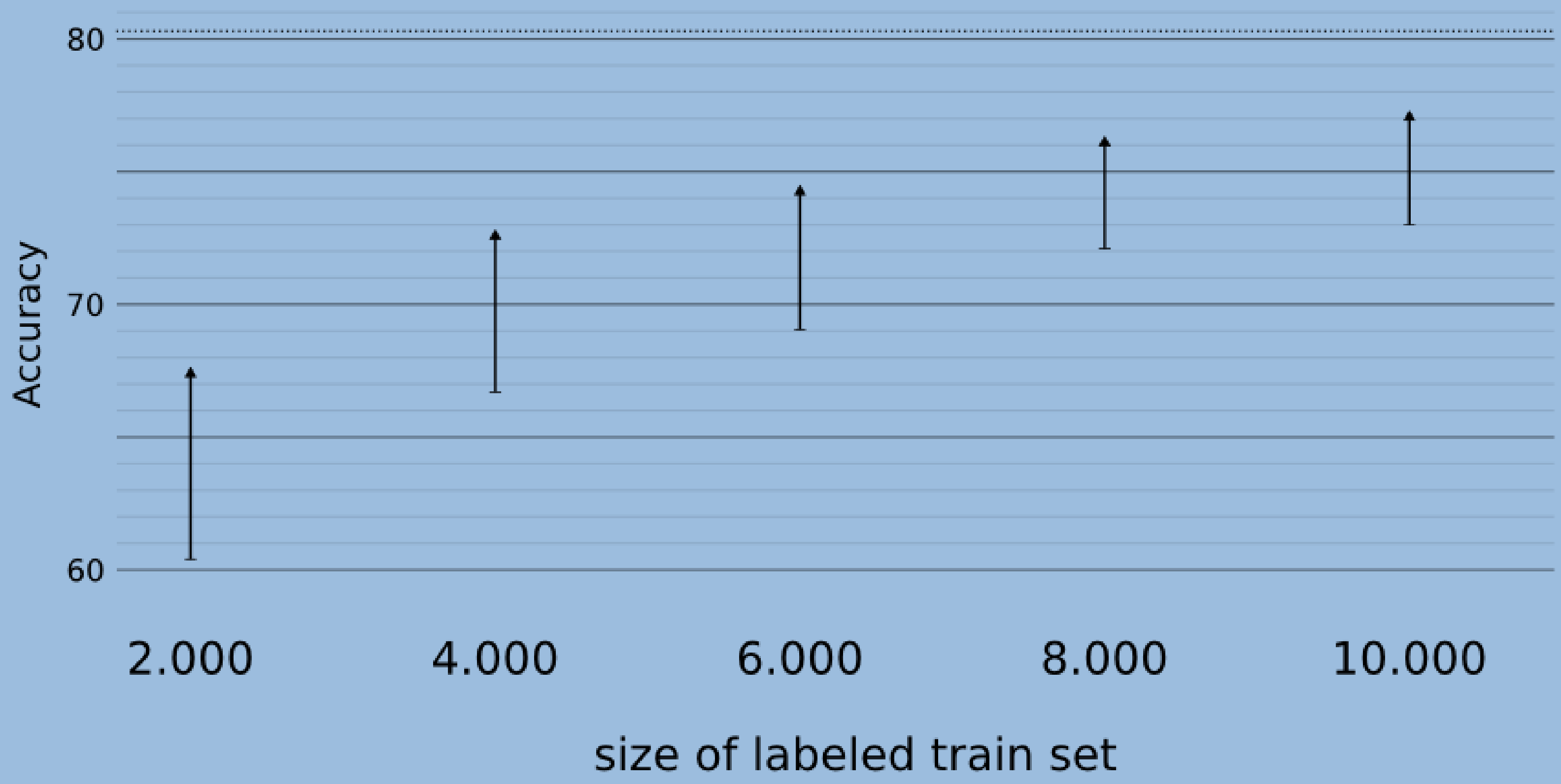
2000 labeled words, with dictionary

UNIVERSITÄT BERN

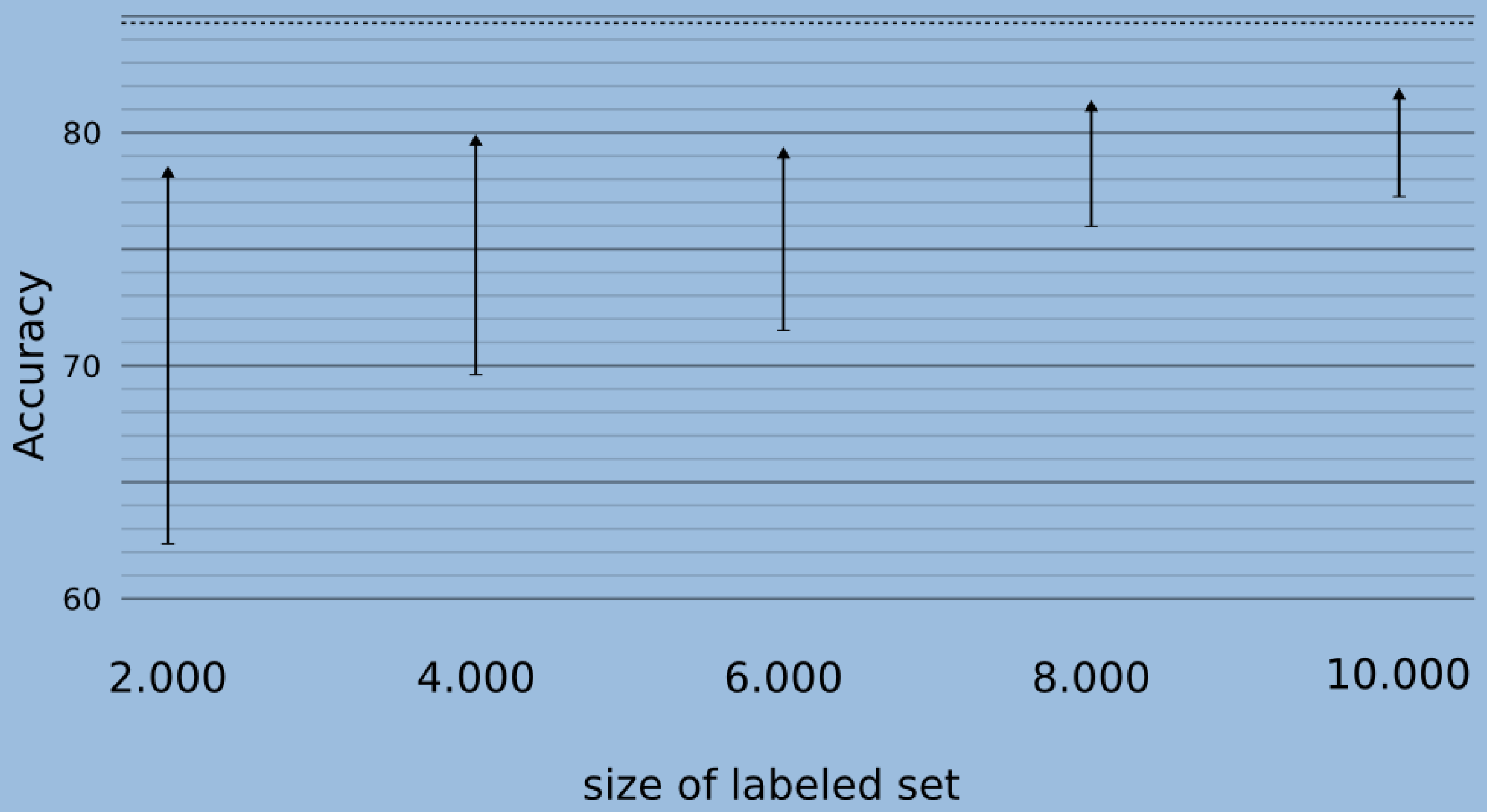
Low Thresholds Retr. Rule



Increase in Accuracy, no dictionary



Increase in Accuracy, with dictionary



b
UNIVERSITÄT
BERN

Results Semi-Supervised Learning

- A significant increase recognition accuracy is possible
- The retraining rules used to create the new training set are crucial for the success
- The optimal retraining rules are hard to estimate beforehand.

- The next steps will include
  - Self-Learning with HMM
  - Co-Learning HMM-NN
  - SSL for text lines