

# Unsupervised Segmentation of Bilingual Text

---

Slav Petrov

[slav@petrovi.de](mailto:slav@petrovi.de)

Fall 2004

# Motivation

- n Given a bilingual text segment all words into two classes.
- n Of interest to linguists:
  - n How do bilingual speakers mix languages?
    - n Depending on the topic?
    - n Depending on part of speech?
- n Of interest to computer scientists:
  - n Unsupervised learning.
- n Of interest to biologists, since it is similar to:
  - n finding the borders between coding and noncoding DNA,
  - n determining the secondary structure of a protein.

# The Data Set

- n Collection of emails written in a mixture of Bulgarian and German.
- n Three different authors (all bilingual).
- n Main language is Bulgarian, but single words, phrases or whole sentences in German.
- n New word creations: German stem + Bulgarian ending.
  
- n From Emails to a Data Set:
  - n Use only body.
  - n Transform to lower case.
  - n Remove all punctuation marks and special characters.
- n Split data set (25,000 words) into:
  - n unlabeled training set (95%),
  - n labeled validation set (2.5%),
  - n labeled test set (2.5%).

# Character based Segmentation (Naive Bayes Classifier)

$$P(c | w, \alpha, \beta) \propto P(c | \alpha) \cdot P(w | c, \beta) = P(c | \alpha) \cdot \prod_{n=1}^N P(a_n | c, \beta)$$

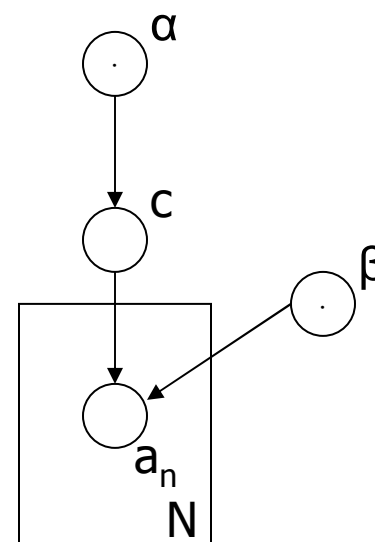
where  $c$  is the language class and  
 $w = a_1 a_2 a_3 \dots a_N$  is a word consisting of characters  $a_n$ .

Assumptions:

- character generation independent of context,
- probability of character is independent of its position in the word,

Learn  $P(c|\alpha)$  and  $P(a_n|c,\beta)$  with EM (not enough labeled data to estimate it directly).

Extend this approach to higher n-gram models (bigrams, trigrams, fourgrams).



# Unsupervised Segmentation

- n Apply Expectation Maximization:
  - n Initialize  $P(c|\alpha)$  and  $P(a_n|c,\beta)$  uniformly (+ noise).
  - n **E-Step:** Label words using current  $P(a_n|c,\beta)$ .
  - n **M-Step:** Estimate  $P(a_n|c,\beta)$  based on current labels.
  - n Iterate until convergence (or for maximum number of iterations).
  
- n Why should this learn to distinguish languages?
  - n Bulgarian (a Slavic language) and German (a Indogermanic language) are fairly different languages.
  - n Relative frequencies of n-grams should be different.

# First Results

- n Evaluate performance on hand labeled test set.
- n Extra class for words that do not influence the performance:
  - n Proper names,
  - n words that are mixtures of both languages (German stem, Bulgarian ending or vice versa).
- n Baseline: randomly choose one class, accuracy 0.5.
- n Model – limitations:
  - n unigrams: 0.8626817447495961
  - n bigrams: 0.7574313408723748
  - n trigrams: 0.9822294022617124
  - n fourgrams: 0.9967689822294022

# Sample In- and Output

## n Input:

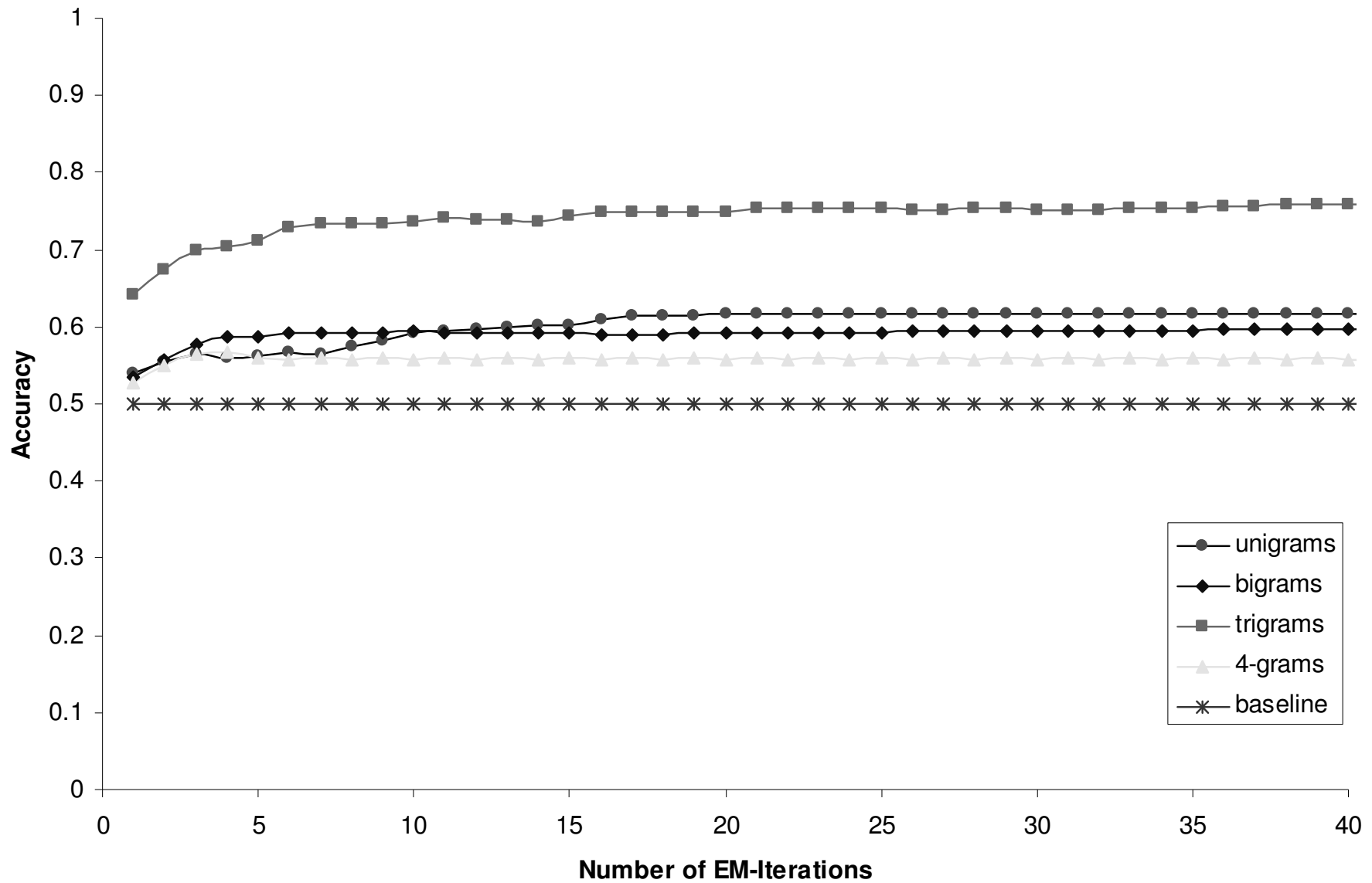
n wchera islogbata besche mnogo interessant maladi kuenstler beshe w rohbau na ubahn reichstag sawsem modern instalazii dosta verrueckt no imasche ot mladite kuenstler koito objasnjawaha ausstellung se kaswasche fraktale wseli sa begriff ot naturwissenschaft sled towa hodihme da jadem w restaurant do brandenburger tor kadeto schroeder s bush hodi kato toi beshe tuk pres mai hubawo napraveno kato literatur haus na ednata stena ima knigi moge da gi chetesch ili da si kupish ako iskasch i ne e chak tolkowa skapo sasega towa e

## n Output:

wchera islogbata besche mnogo interessant maladi kuenstler beshe w rohbau na ubahn reichstag sawsem modern instalazii dosta verrueckt no imasche ot mladite kuenstler koito objasnjawaha ausstellung se kaswasche fraktale wseli sa begriff ot naturwissenschaft sled towa hodihme da jadem w restaurant do brandenburger tor kadeto schroeder s bush hodi kato toi beshe tuk pres mai hubawo napraveno kato literatur haus na ednata stena ima knigi moge da gi chetesch ili da si kupish ako iskasch i ne e chak tolkowa skapo sasega towa e

green: correct Bulgarian, blue: correct German, red: incorrect

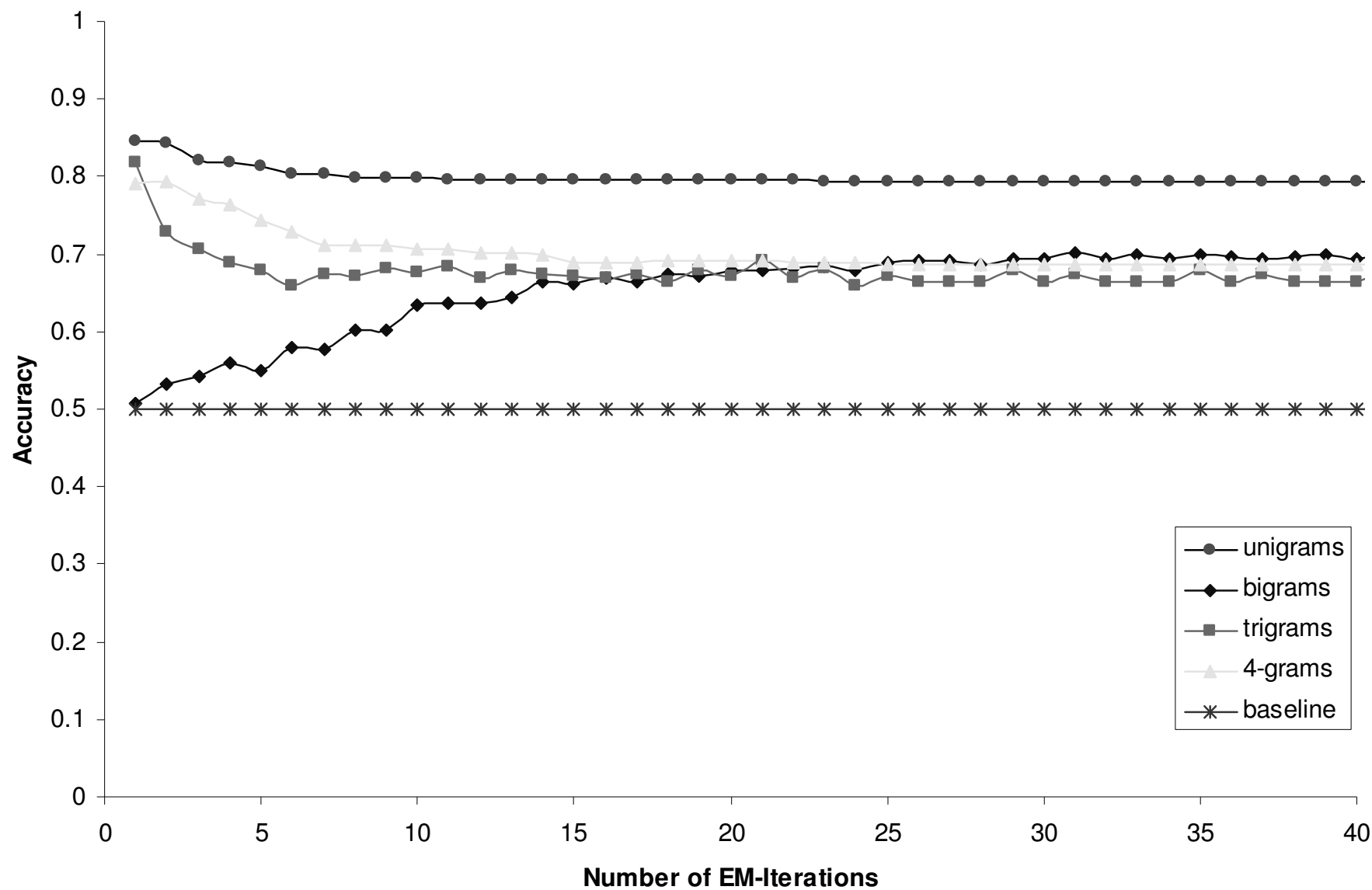
# Unsupervised Segmentation Performance



# Semi-Supervised Segmentation

- n Use the validation set to roughly estimate  $P(c|\alpha)$  and  $P(a_n|c,\beta)$ .
- n Use these values instead of the random initialization.

# Semi-Supervised Segmentation Performance (EM)



# Deterministic Annealing 1

## n Observations:

- n Higher likelihood usually correlates with higher accuracy,
- n BUT: performance highly dependent on initialization of EM (local maxima).
- n => Search error

## n Solution:

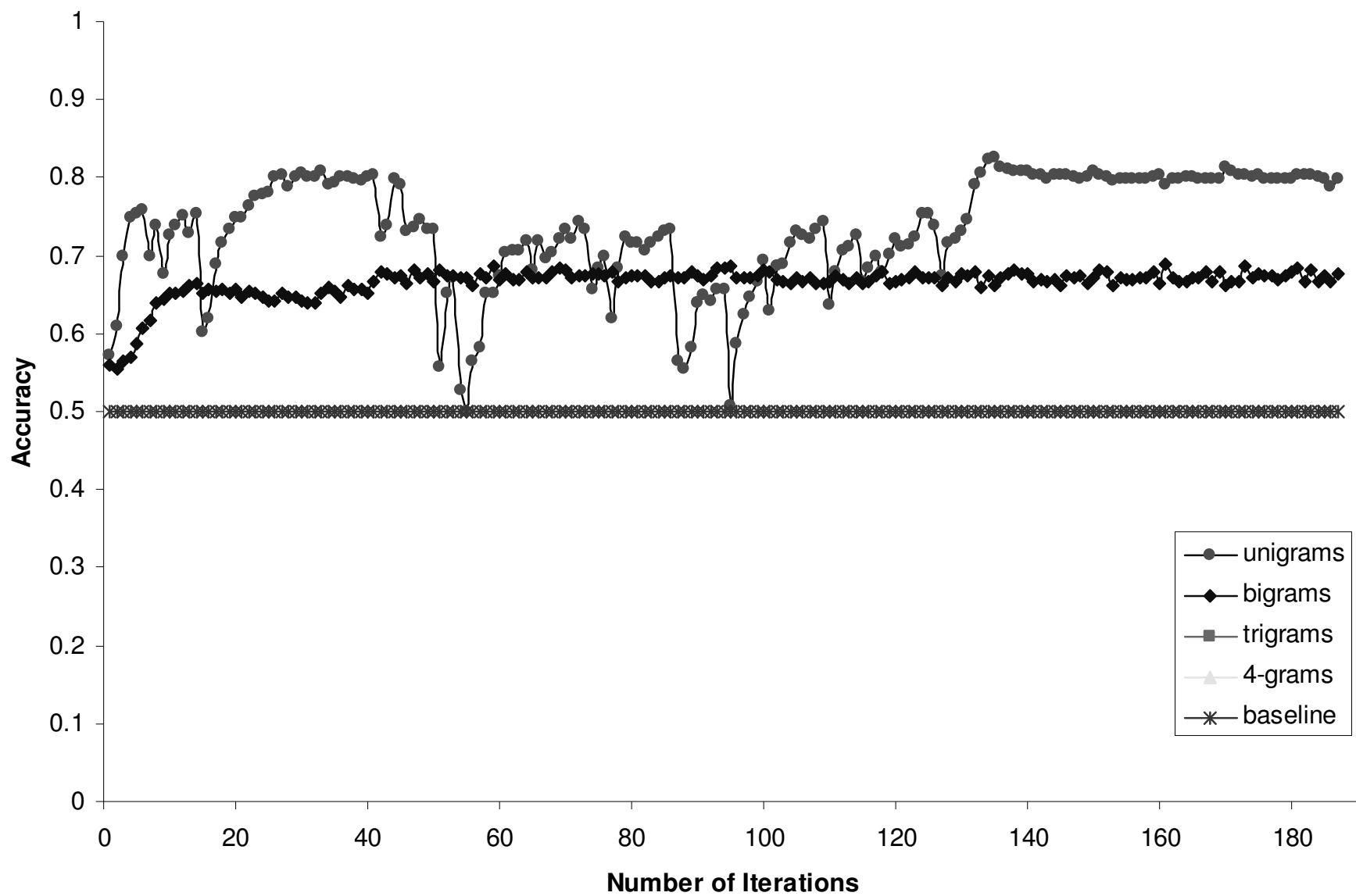
- n Start with an easy concave function.
- n Morph function gradually into desired non-concave likelihood function while maintaining a local maximum.

# Deterministic Annealing 2

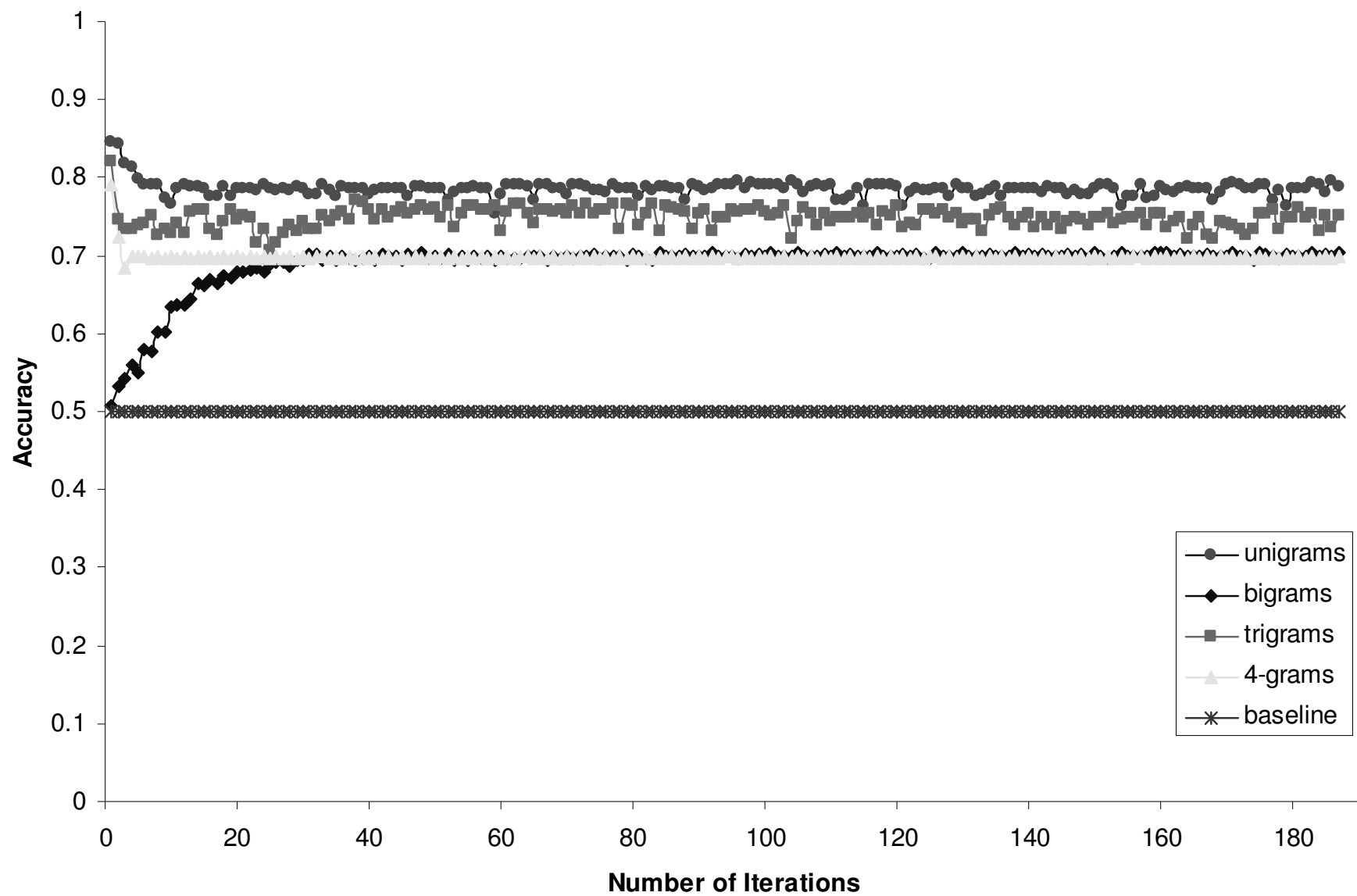
- n Use EM in an inner loop.
- n Add parameter  $\gamma$ 
  - n start with  $\gamma = 0.1$
  - n increase  $\gamma$  by a factor of 1.2 after each iteration of DA as long as  $\gamma \leq 1.0$
- n Change E-Step to:

$$P(c | w, \alpha, \beta) \propto P(c | \alpha)^\gamma \cdot P(w | c, \beta)^\gamma$$

# Unsupervised Segmentation Performance (DA)



# Semi-Supervised Segmentation Performance (DA)



# Conclusions & Future Work

- n Classifier sometimes gets “lucky”.
- n Training set and especially test set are too small.
- n Best performance when  $P(c|\alpha)$  and  $P(a_n|c,\beta)$  are estimated from the validation set.
- n Unsupervised learning decreases the accuracy!
  
- n More data necessary.
- n Implement classifier on word level:
  - n Hidden Markov Model or
  - n Latent Dirichlet Allocation (languages instead of topics).

# Future Work

---

- n More data necessary.
- n Implement classifier on word level:
  - n Hidden Markov Model or
  - n Latent Dirichlet Allocation (languages instead of topics).

Thank you very much for your  
attention.

---

For more information:

<http://www.eecs.berkeley.edu/~petrov>



```
ERROR: undefined
OFFENDING COMMAND:

STACK:
```