Practical Applications for Language Models

Dietrich Klakow
Spoken Language Systems
Saarland University

Overview

- Vocabularies for morphologically rich languages
- Selecting task specific sub-corpus
- LM adaptation using tiny adaptation corpus

Inventing new Words

Vocabularies for morphologically rich languages

Tamil

Training corpus: 103651 tokens

Test corpus: 94664 tokens

Vocabulary size: 16209 (=types in training corpus)

Most frequent words: **ஆஹ்** (oh!), **உம்ம்** (hmmm), **சரி** (okay) , **என்ன** (what?) , **நான்** (I), **அது** (that), **நீ** (you)

Problem: 13% OOV rate

Frequent OOVs: **லவ்** (love), **ஆனி** (a month), **சரிணா** (okay?), **குணா** (name), **என்னாங்க** (what) , **வாடக** (rent)

Idea:

- Decompose words into smaller units
- Train language model on the smaller units
- Sample from it
- Create extended vocabulary from artificial corpus

Decompose words

- Used morfessor on flat vocabulary

<s> அக்கா க்கா </s>
<s> அக்கா க்கு </s>
<s> அக்கா டா </s>
…

Frequent units:

ல 592
ம் 491
னு 419
ன் 409
ங்க 370
மா 366
டா 356
க்கு 355
து 336

Language Modeling and Sampling

- Train trigram LM from fragments
- Sample from the LM

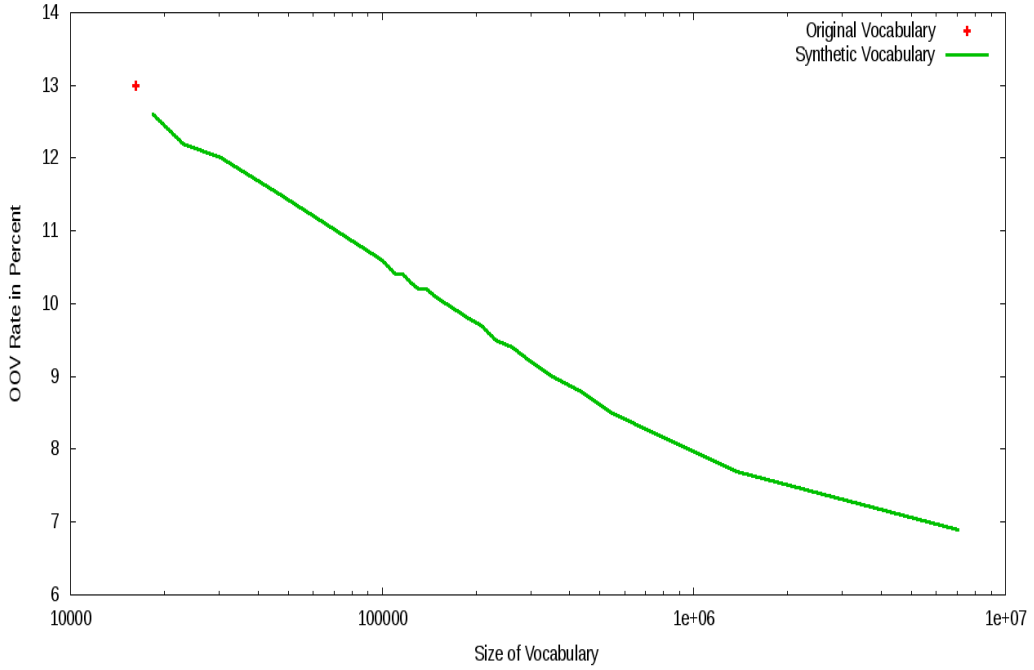
Language Modeling and Sampling

- Train trigram LM from fragments
- Sample from the LM
- Examples:

போயிருந்த	“went”
வரி	“tax”
பெரு	“big”
வந்துர்	*garbage*
குடு	“give”
றான்	*garbage*
மாச்சு	*garbage*
இல்லியா	“not there?”
கௌம்பு	“go away”

Language Modeling and Sampling

- Train trigram LM from fragments
- Sample from the LM



Next steps

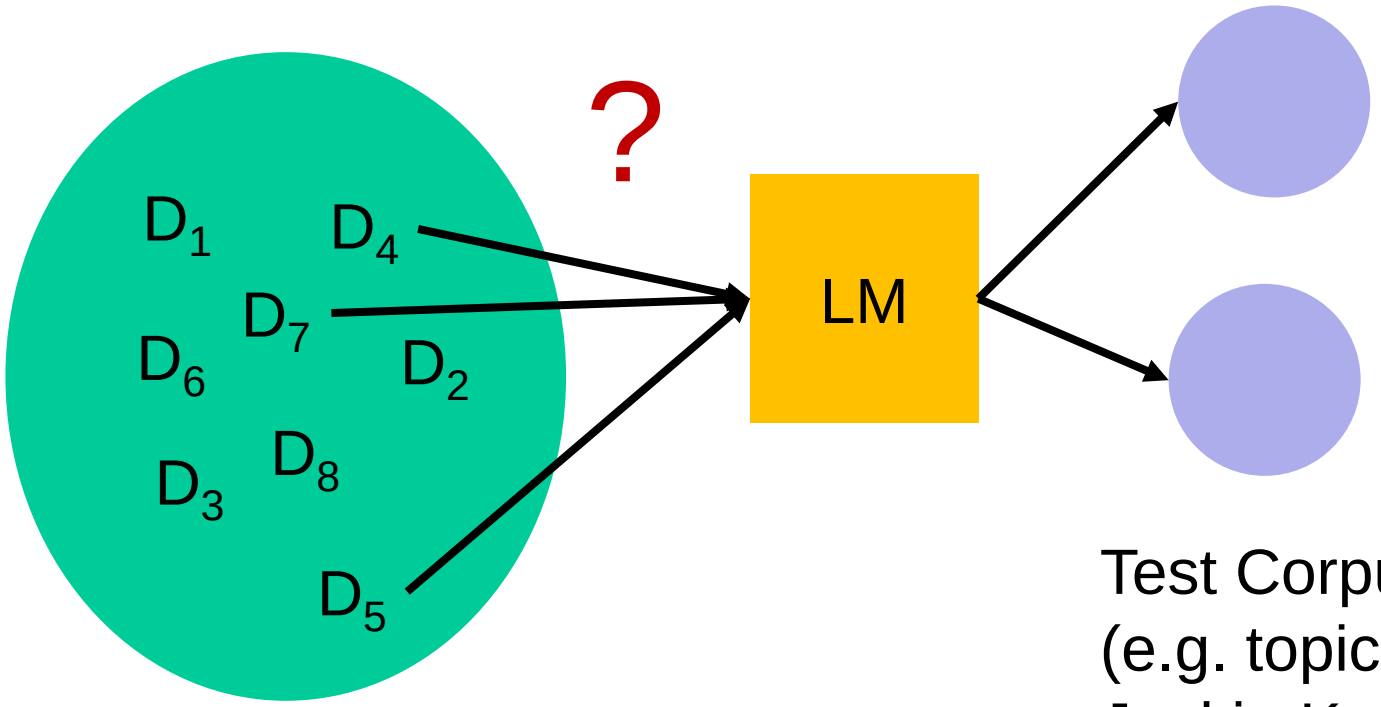
- Use different language models
- Try other languages
- Use methods from Jilles for segmentation

Selecting Documents from a Background Corpus

The Task

Background Corpus (e.g. North American News Text)

Target Corpus (e.g. topic Jackie Kennedy)



Test Corpus (e.g. topic Jackie Kennedy)

Approach

Measure change in likelihood when document is omitted from training corpus:

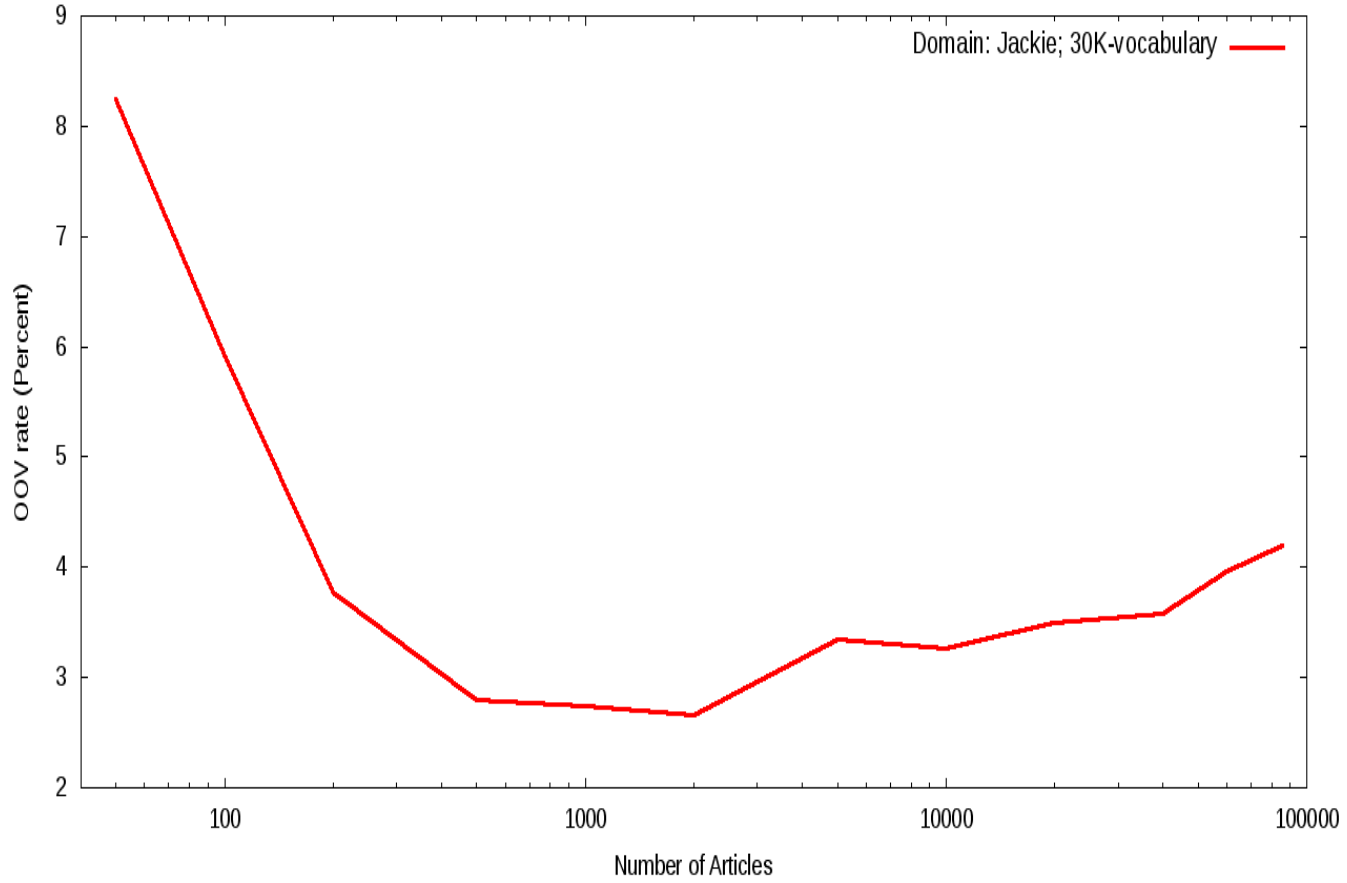
$$\Delta F_i = \sum_w N_{\text{Target}}(w) \log \frac{P(w)}{P_{D_i}(w)}$$

$N_{\text{Target}}(w)$: Unigram counts on target corpus

$P(w)$: Unigram LM on complete training corpus

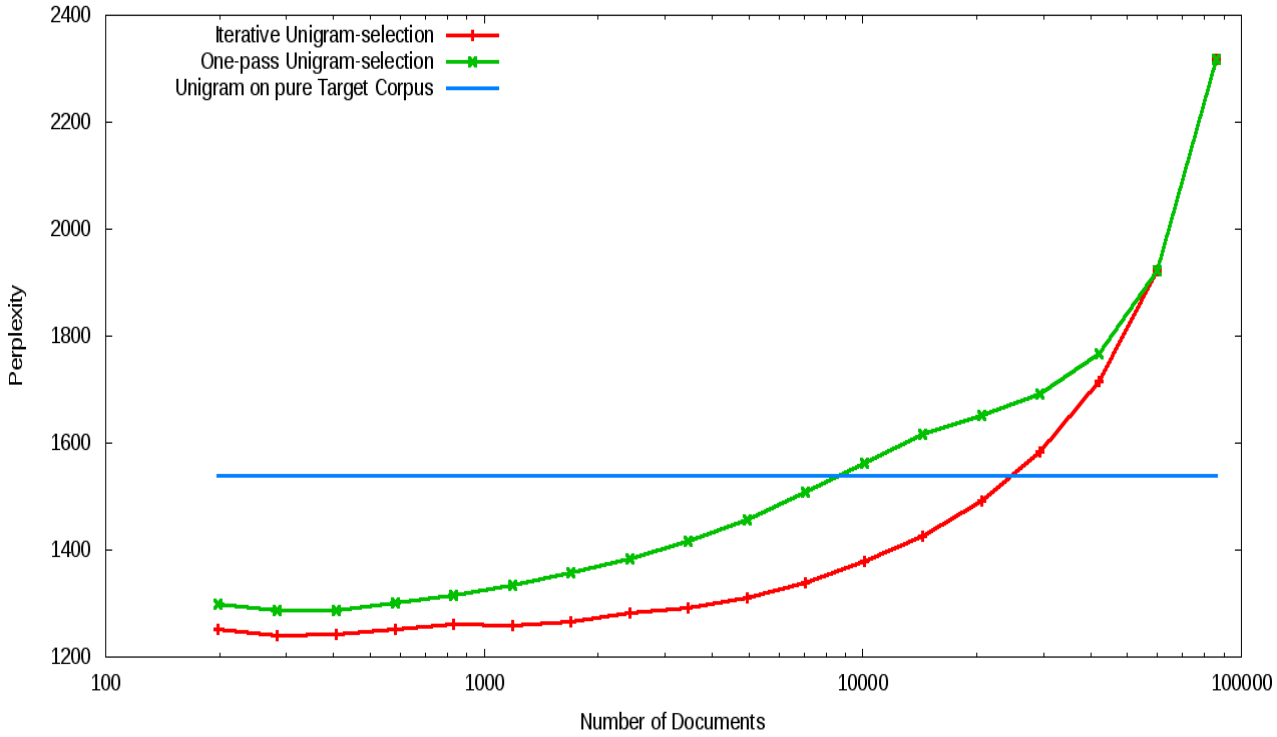
$P_{D_i}(w)$: Unigram LM on complete training corpus when i-th document is omitted

OOV Rate



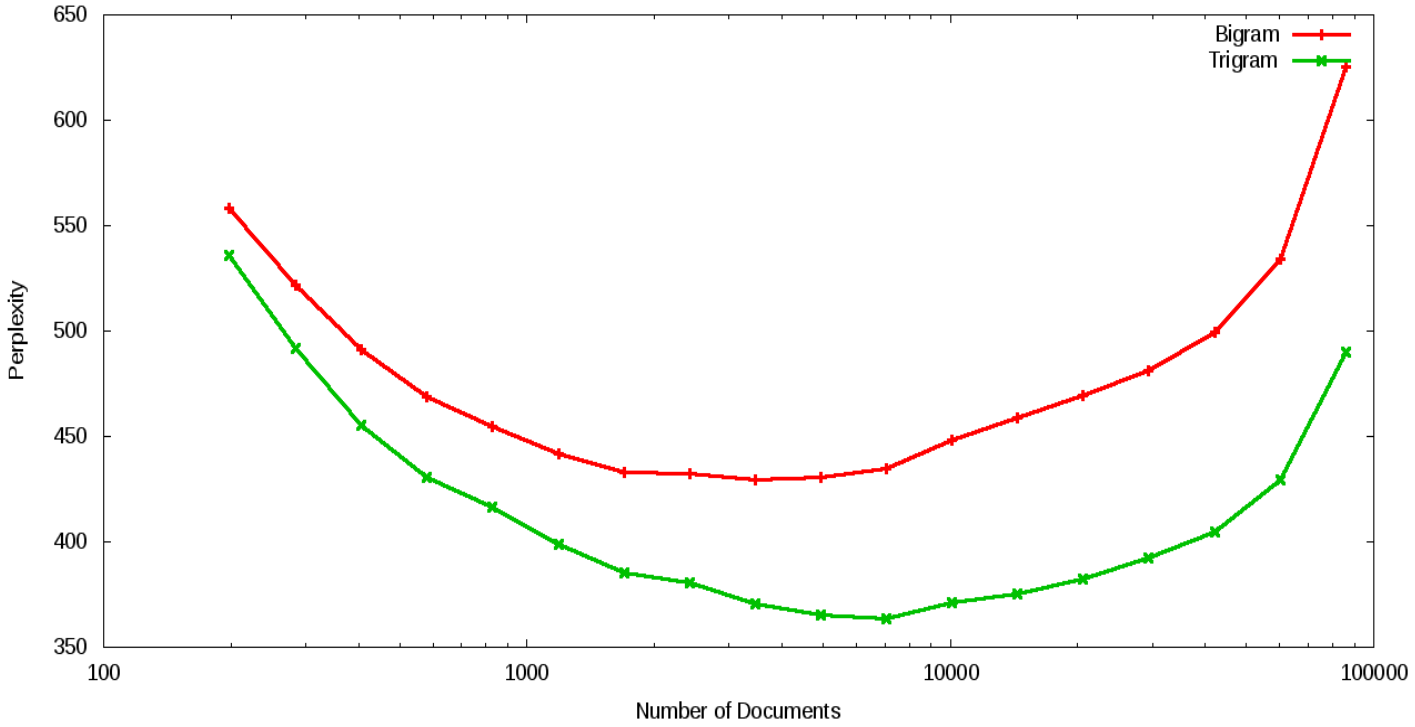
Removing documents reduces OOV rate

Unigram Perplexities



Selection from background better than target corpus

Bi- and Trigram Perplexities



Selection helps bigrams and trigrams

Summary

Mercer's famous comment:
“There is no data like more data”

Sometimes having the right data is important

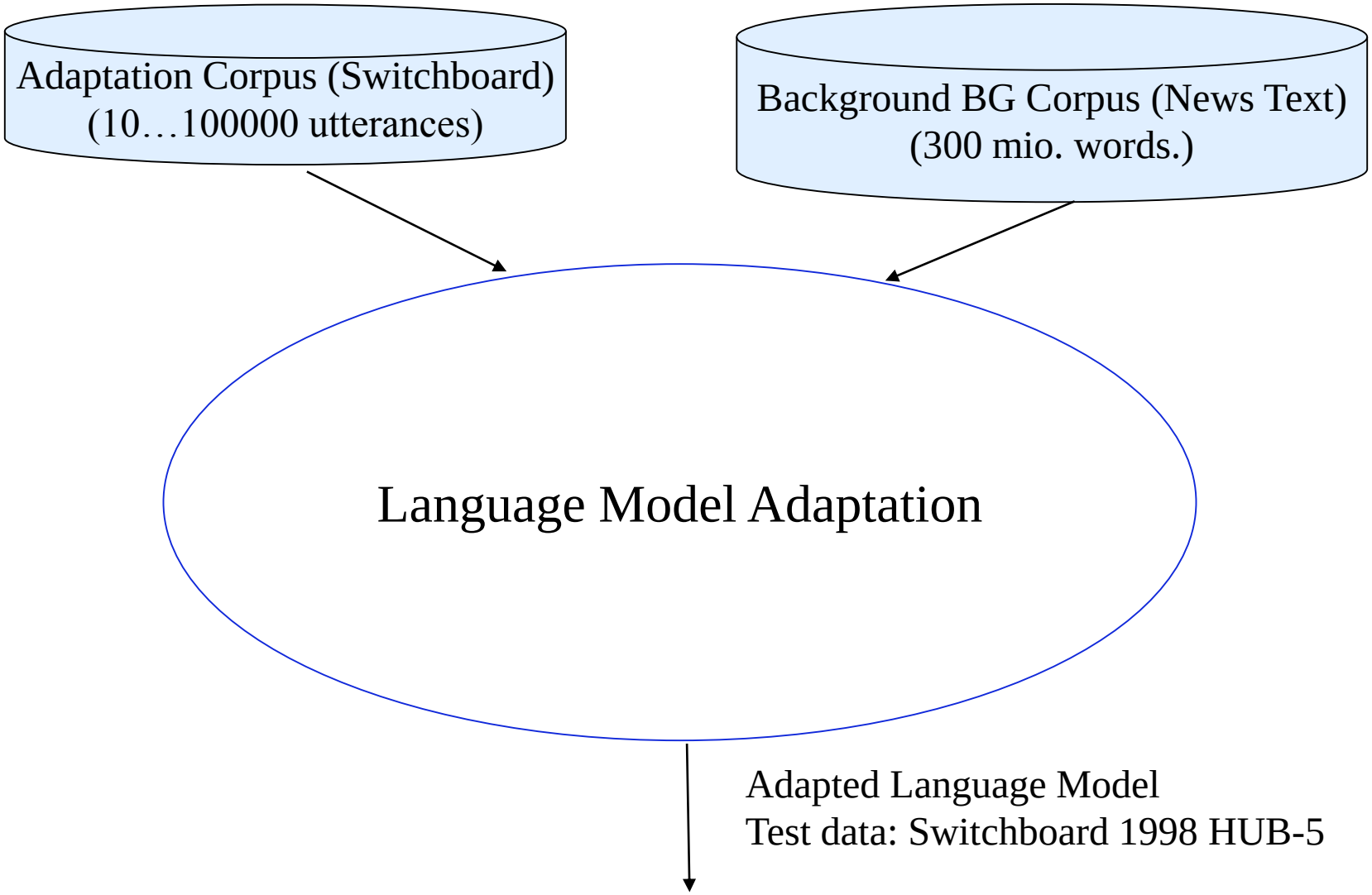
Language Model Adaptation for Tiny Adaptation Corpora

Task

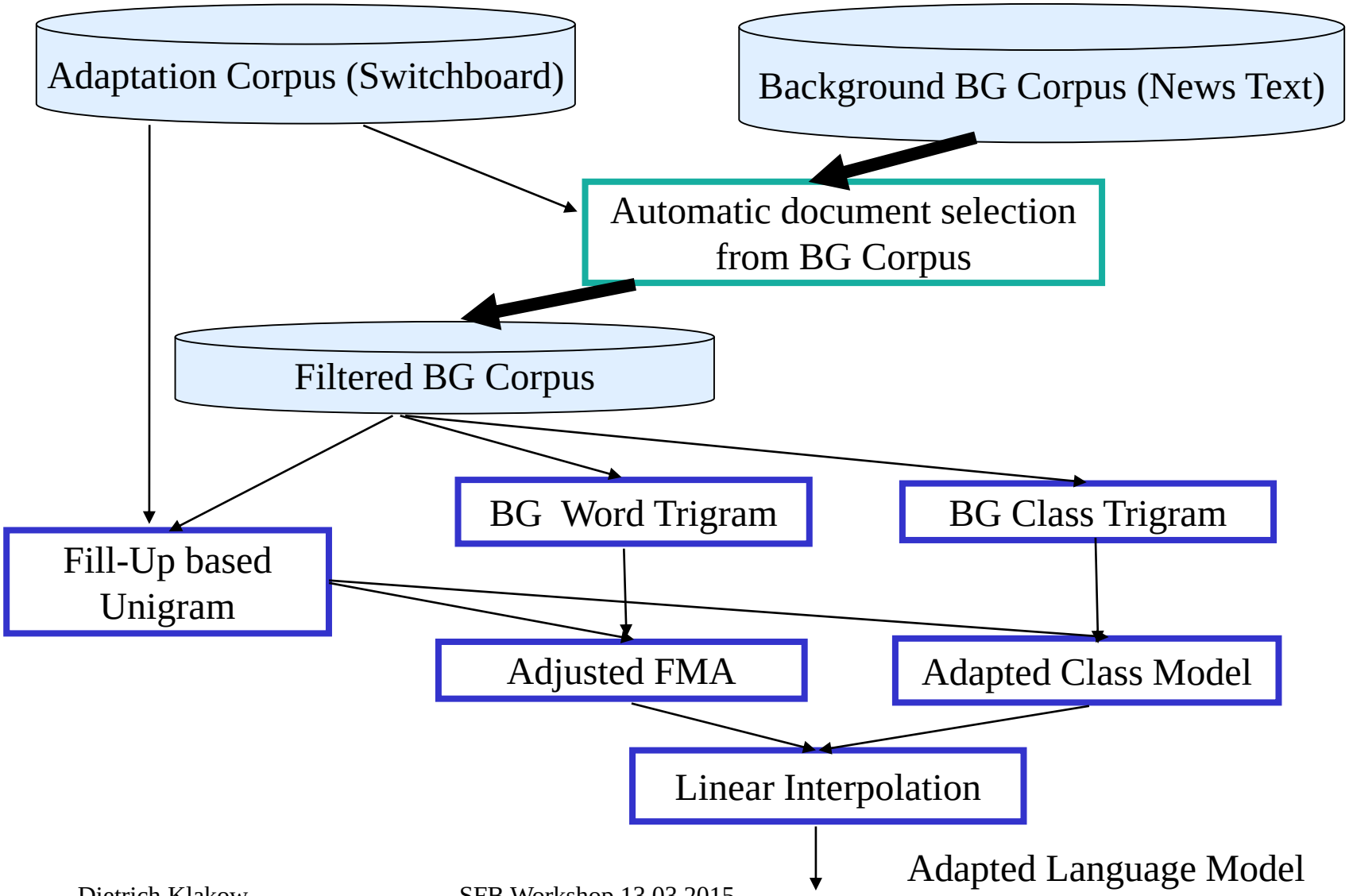
- Adapt a language model using as little adaptation data as possible (e.g. 10 utterances)
- Example:

1. A: okay
2. B: hi
3. A: hi
4. B: um yeah i would like to talk about how you dress for work and and um
5. A: well i work in uh corporate control so we have to dress kind of nice
6. B: uh-huh
7. A: and in the summer just dresses we can't even well we're not even
8. B: and is
9. A: it really doesn't vary that much from season to season since the office is
10. B: right right is there is there um a- is there a like a code of dress

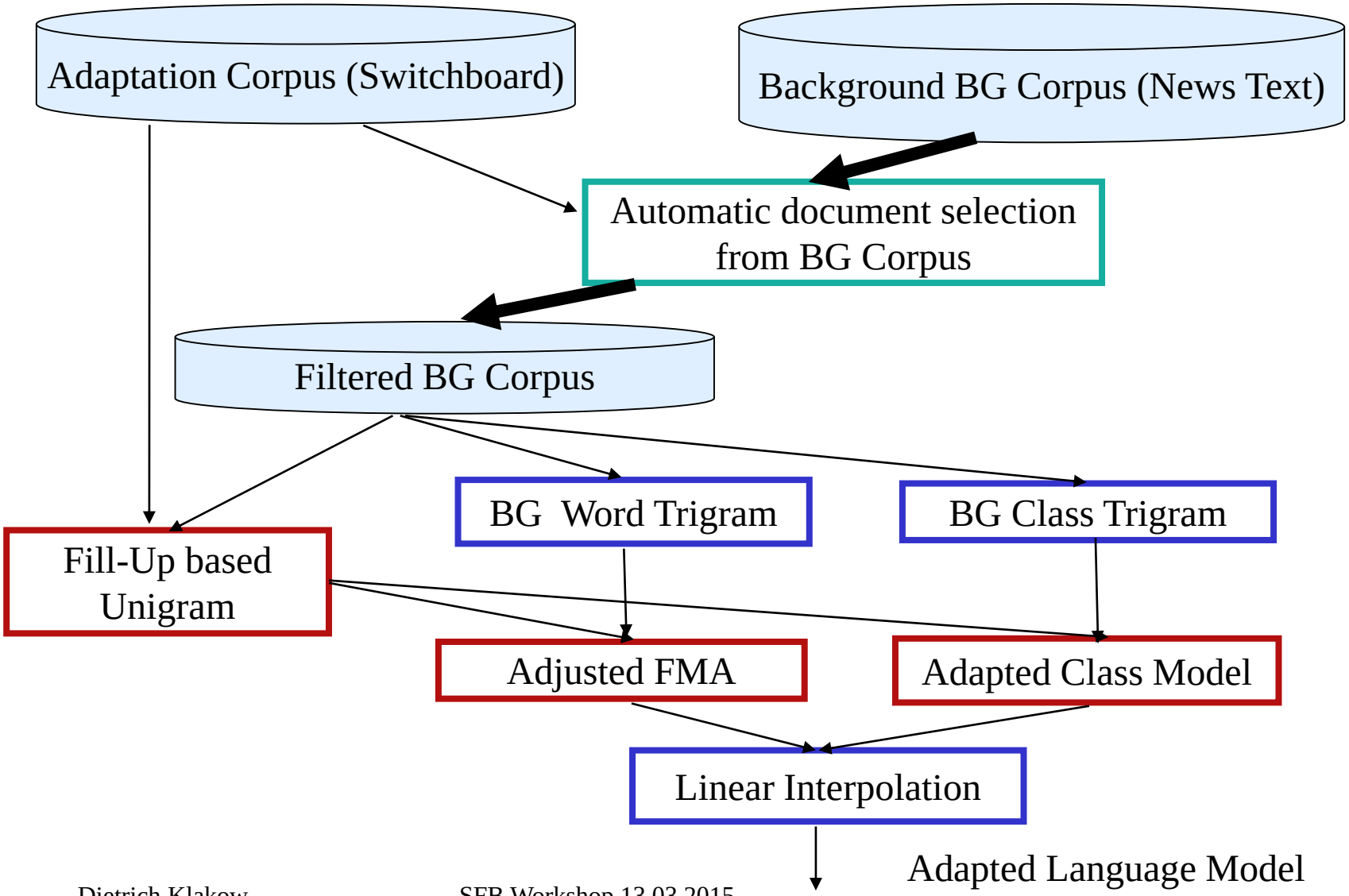
Task


Adaptation Corpus (Switchboard)
(10…100000 utterances)
Background BG Corpus (News Text)
(300 mio. words.)
Language Model Adaptation
Adapted Language Model
Test data: Switchboard 1998 HUB-5

Overview of the System



Overview of the System



Estimating good adapted Unigrams

- Using Fill-Up

$$P_{adap}(w) = \begin{cases} \dfrac{N_{Adap}(w) - d}{N_{Adap}} + \alpha P_{BG}(w) & \text{if } N_{Adap}(w) > 0 \\ \alpha P_{BG}(w) & \text{else} \end{cases}$$

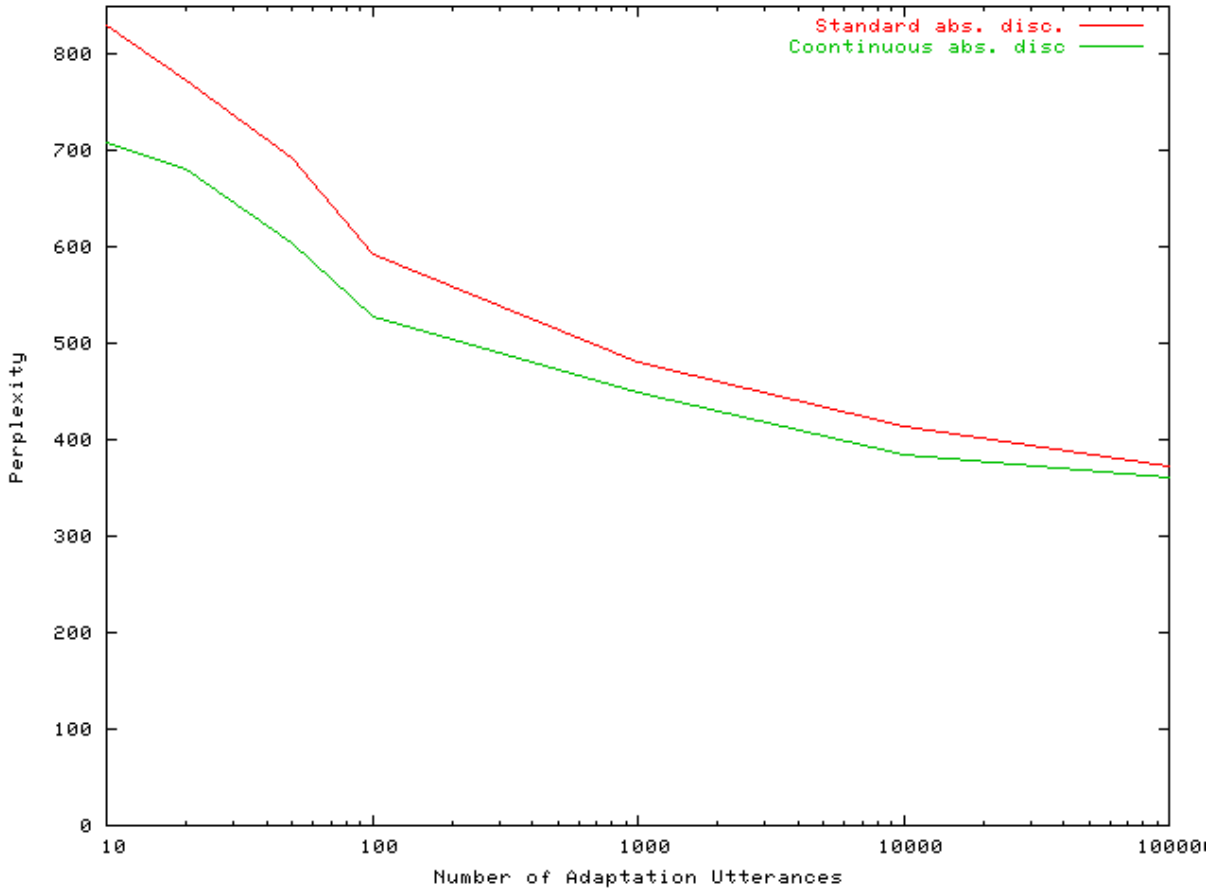
Besling at al. 1995

- Make discounting parameter depend continuously on count

$$d(N) = \frac{d_0 + s(N-1)}{1 + g(N-1)}$$

d_0, s and g are parameters tuned of development set

Effect of Count Continuous Discounting on Fill-Up



Improves unigrams in particular for very small adaptation corpora

PP_{BG}=1830

Adjusted FMA

FMA:

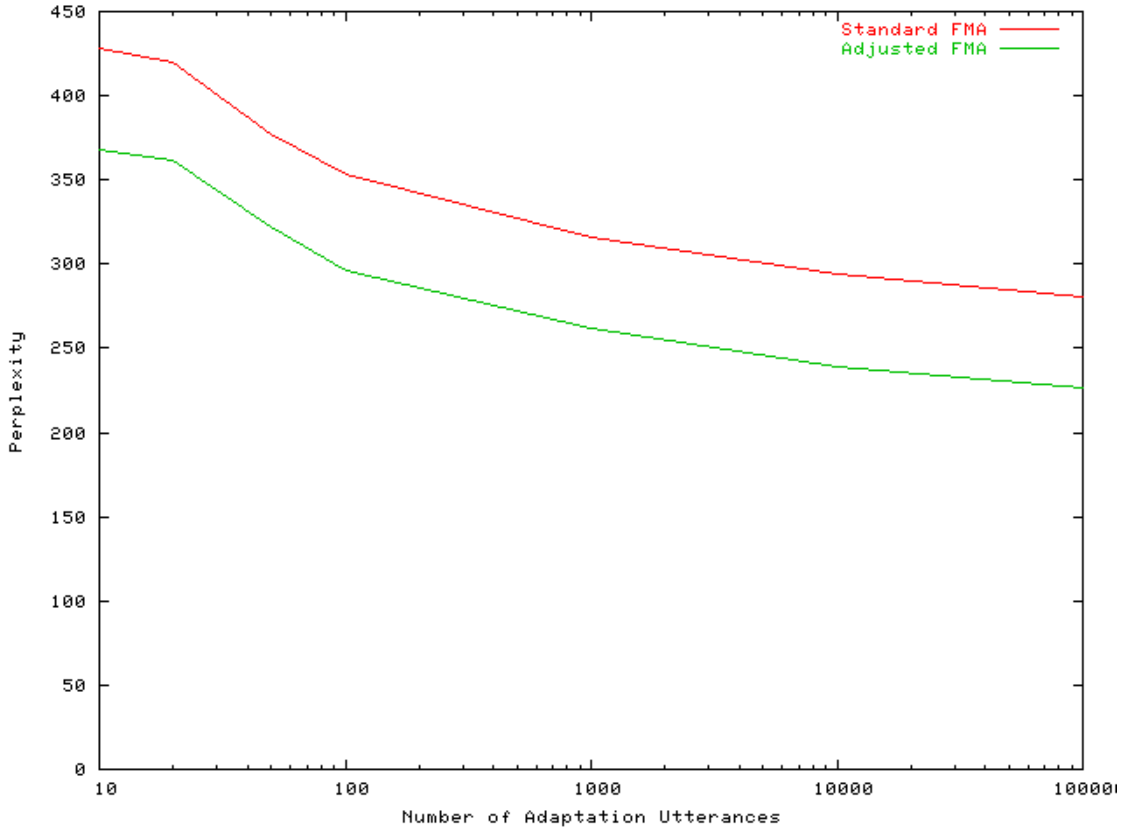
$$P(w \mid h) = \frac{1}{Z(h)} \left(\frac{P_{Adap}(w)}{P_{BG}(w)} \right)^{\beta} P_{BG}(w \mid h)$$

Kneser at al. 1997

Adjusted FMA

$$P(w \mid h) = \frac{1}{Z(h)} \left(\frac{P_{BG}(w \mid h)}{P_{BG}(w)} \right)^{\beta} P_{Adap}(w)$$

Compare Standard and Adjusted FMA



Additional consistent improvement

Adaptation of Class Models

Normal Class Model

$$P(w|h) = P(w|c)P(c|C(h))$$

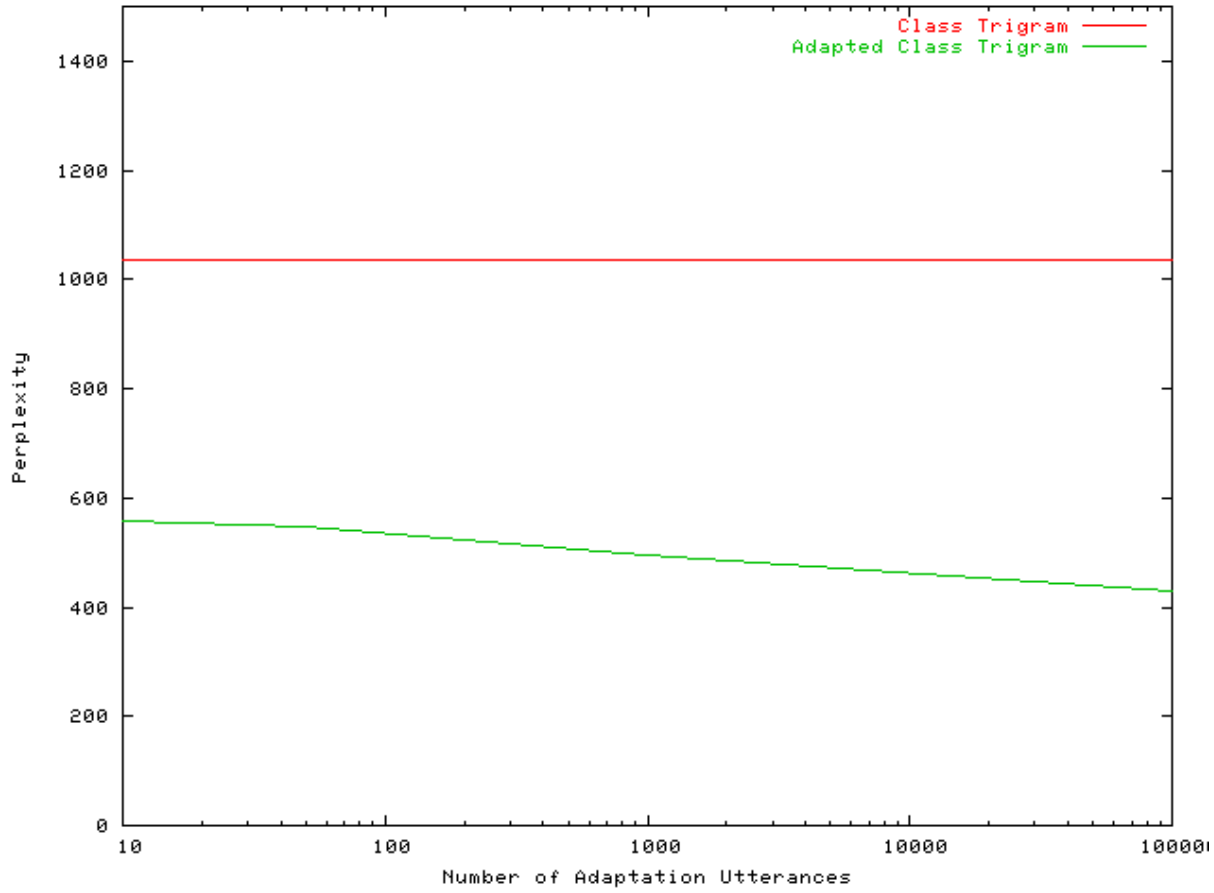
Emission model Class prediction

Idea: *adapt the emission model*

Adapted Class Model

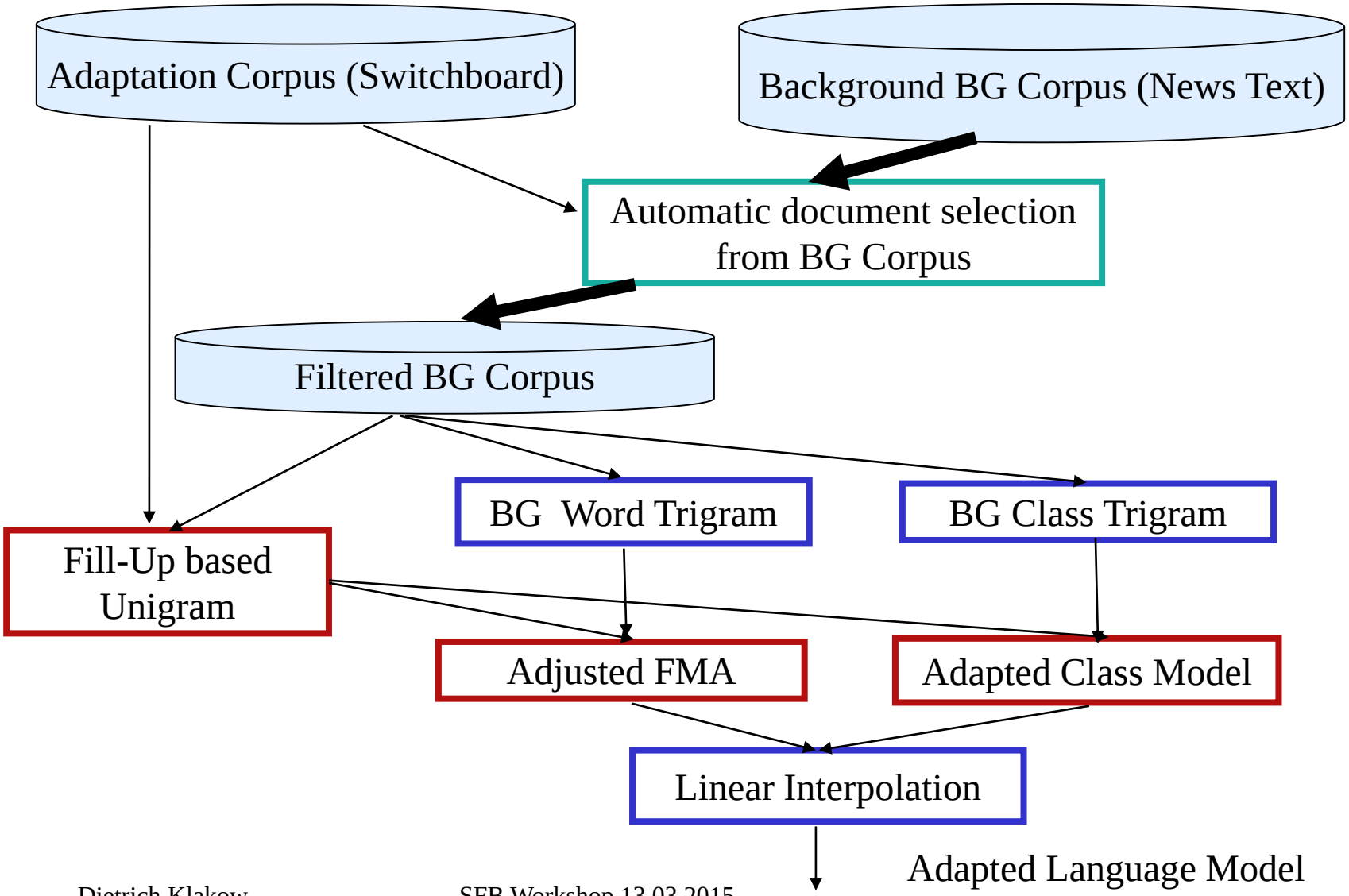
$$P_{Adap}(w|h) = P_{Adap}(w|c)P_{BG}(c|C(h))$$

Results for Adapted Class Model

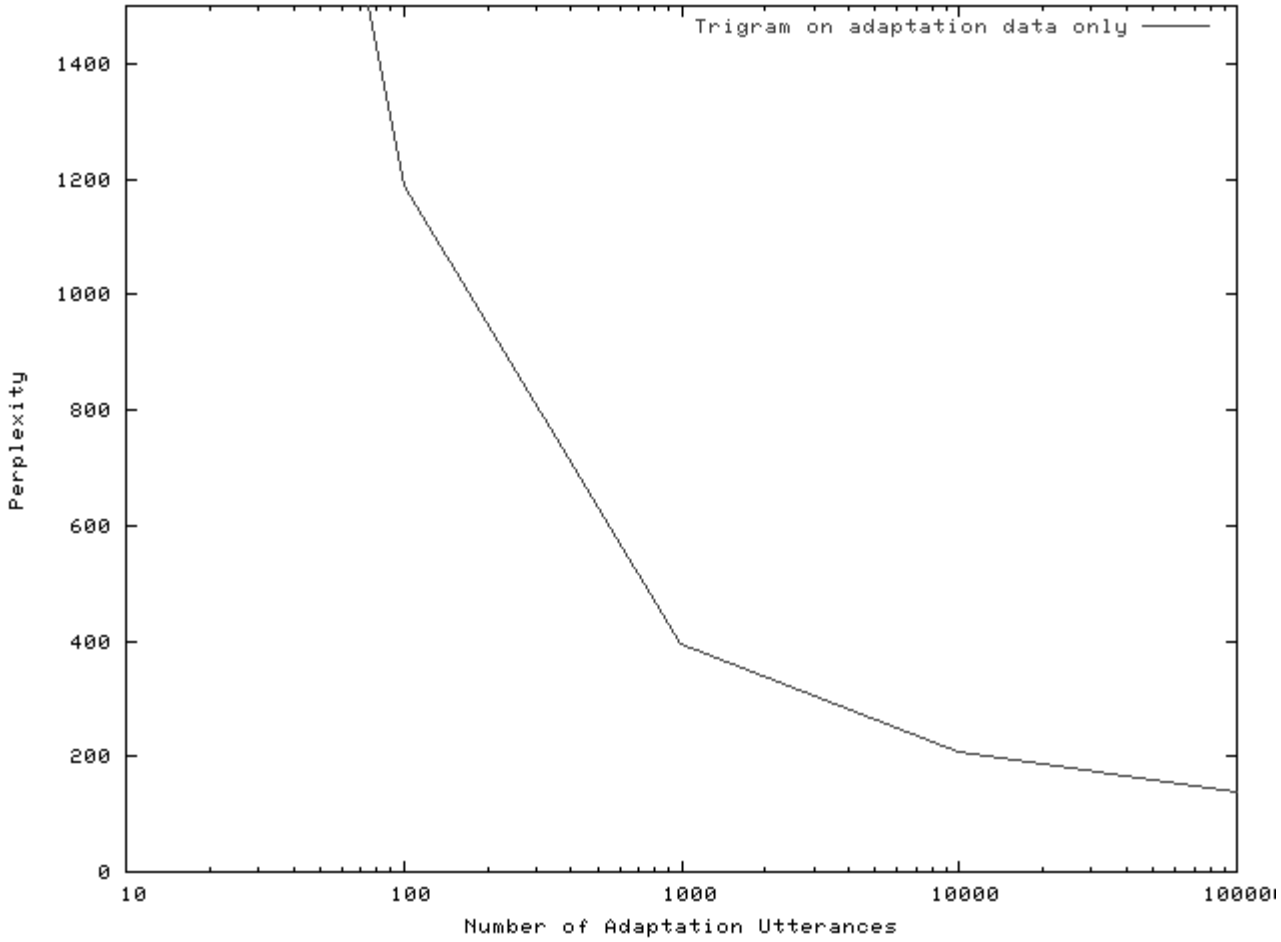


Always a reduction of perplexity by a factor of two!

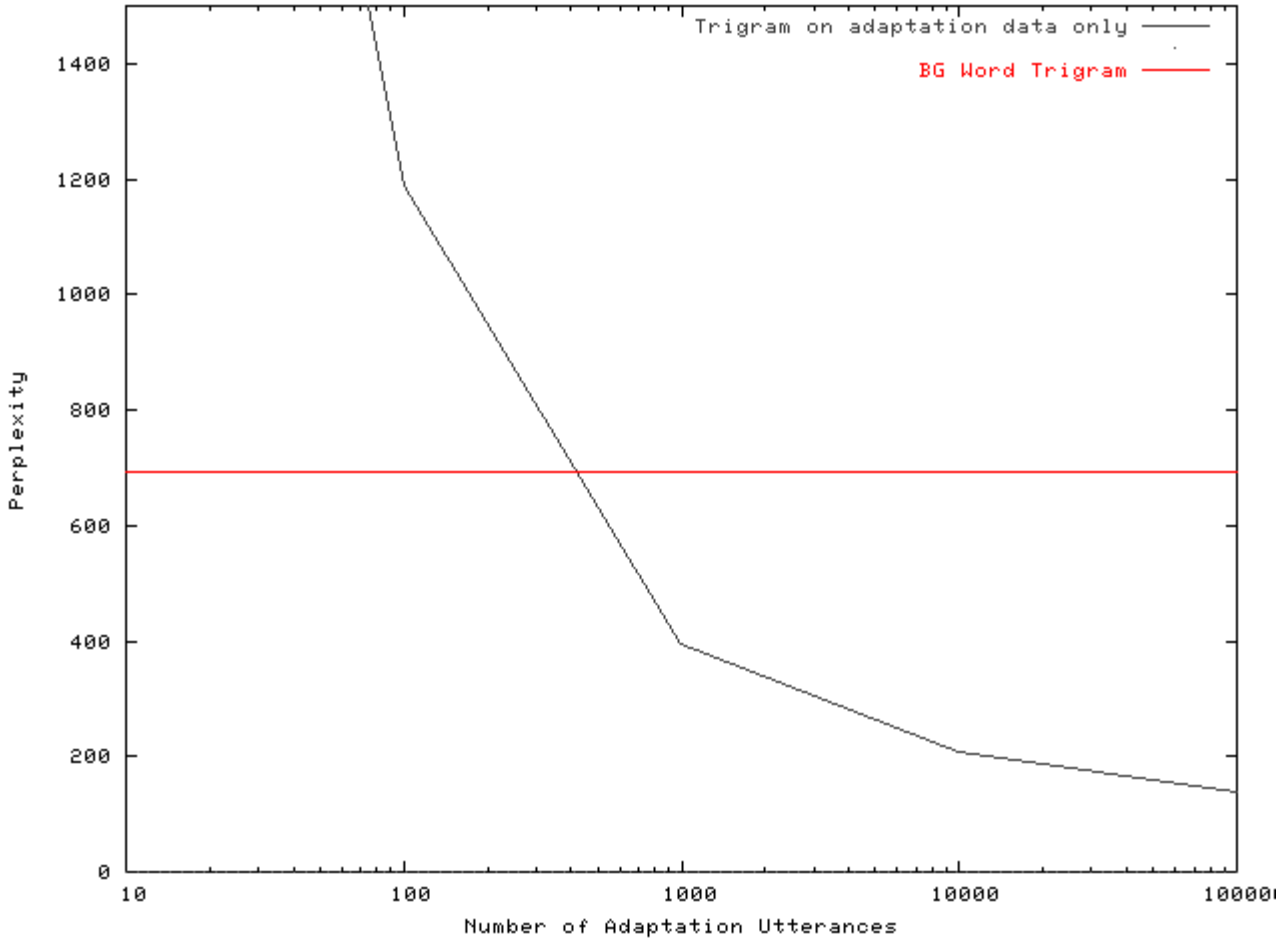
Overview of the System



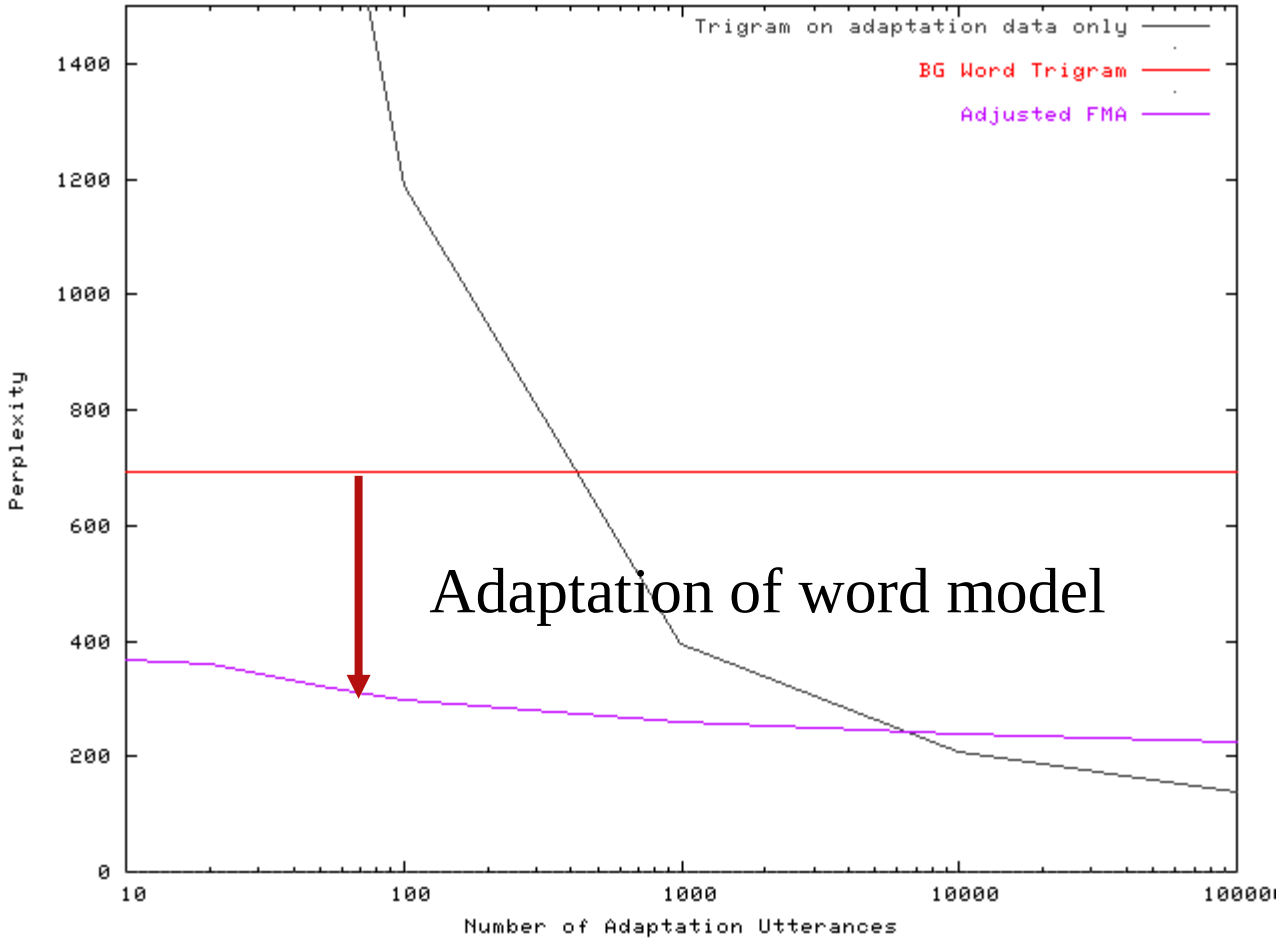
Comparison on Trigram Models



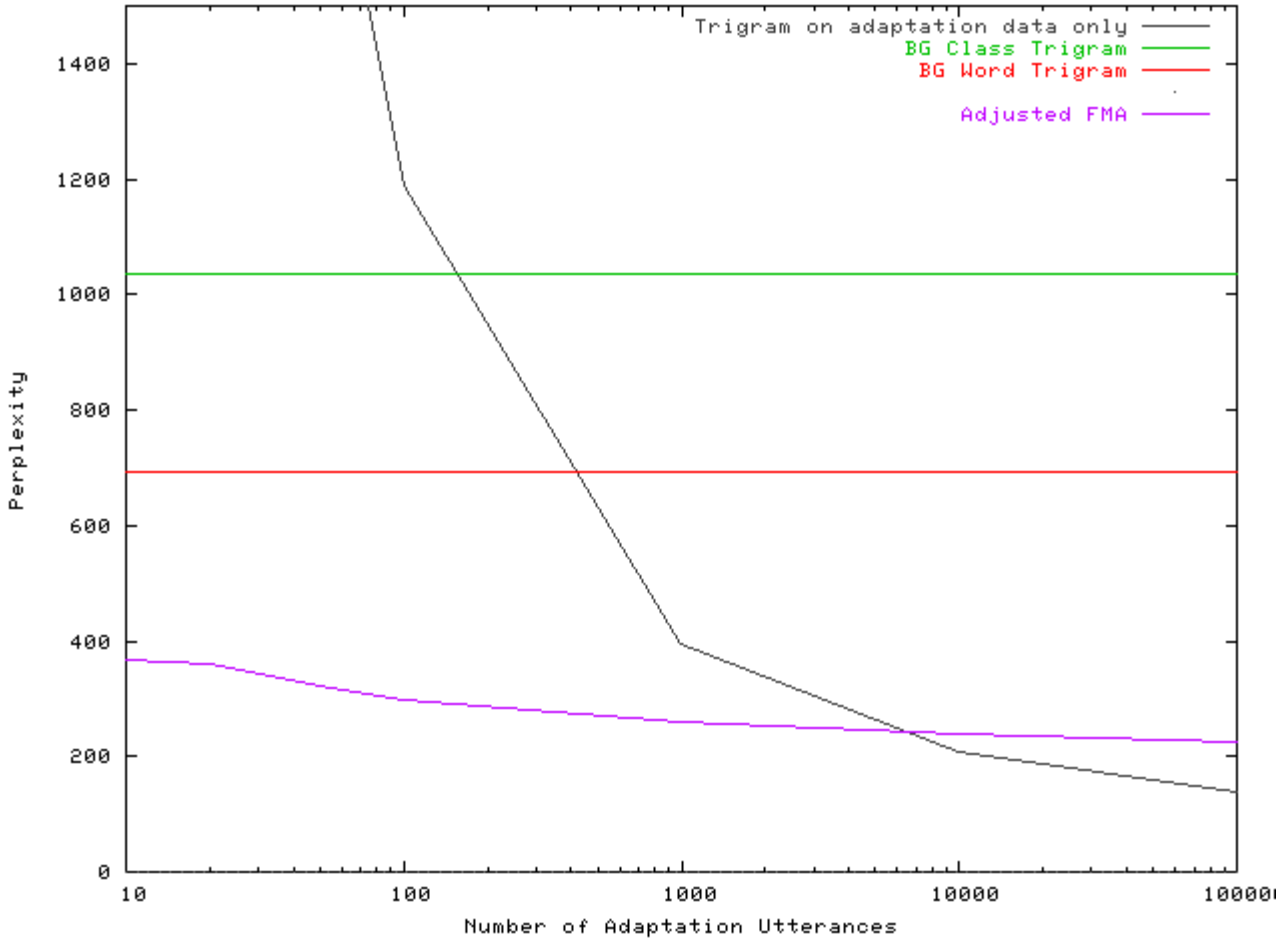
Comparison on Trigram Models



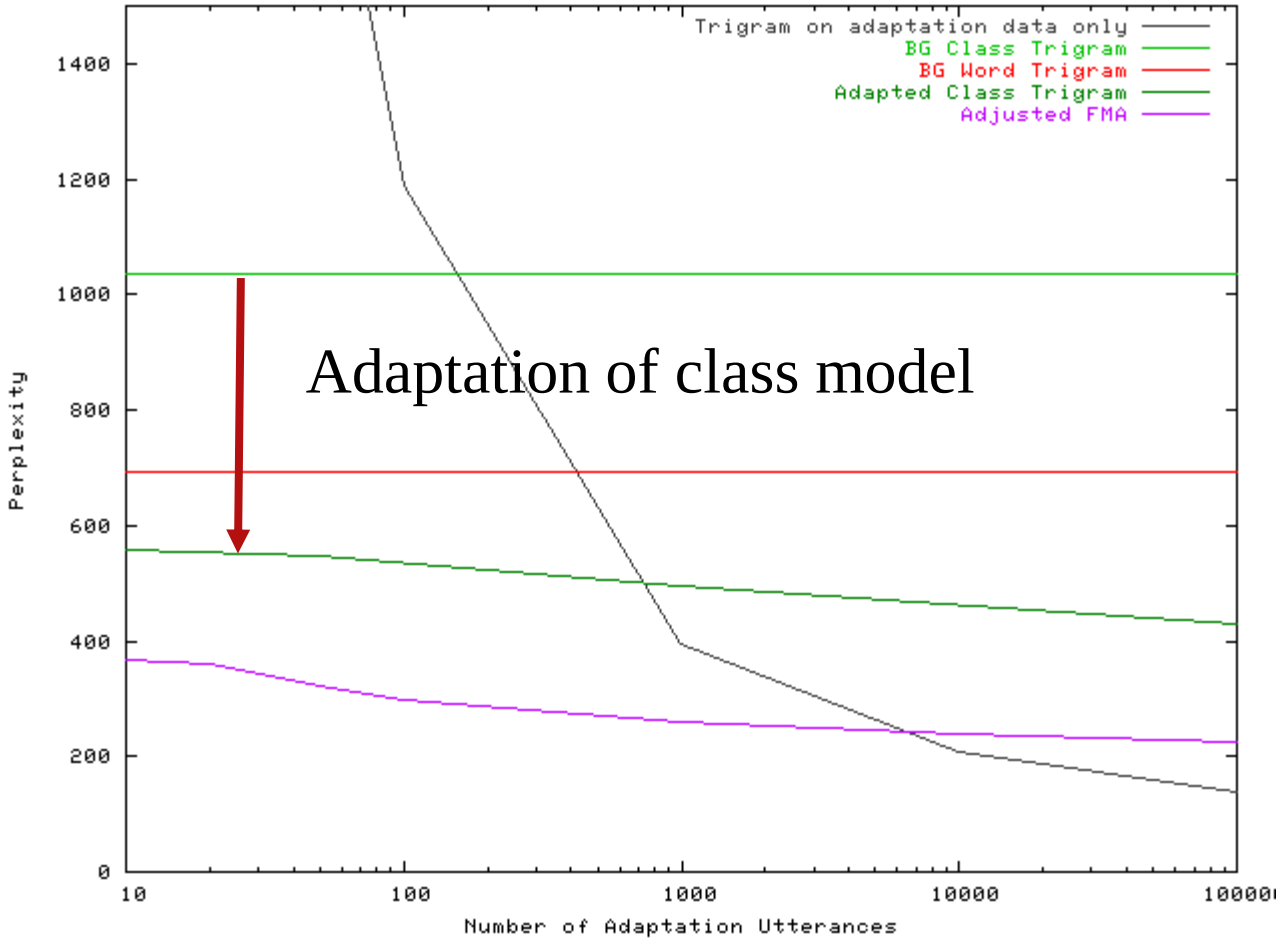
Comparison on Trigram Models



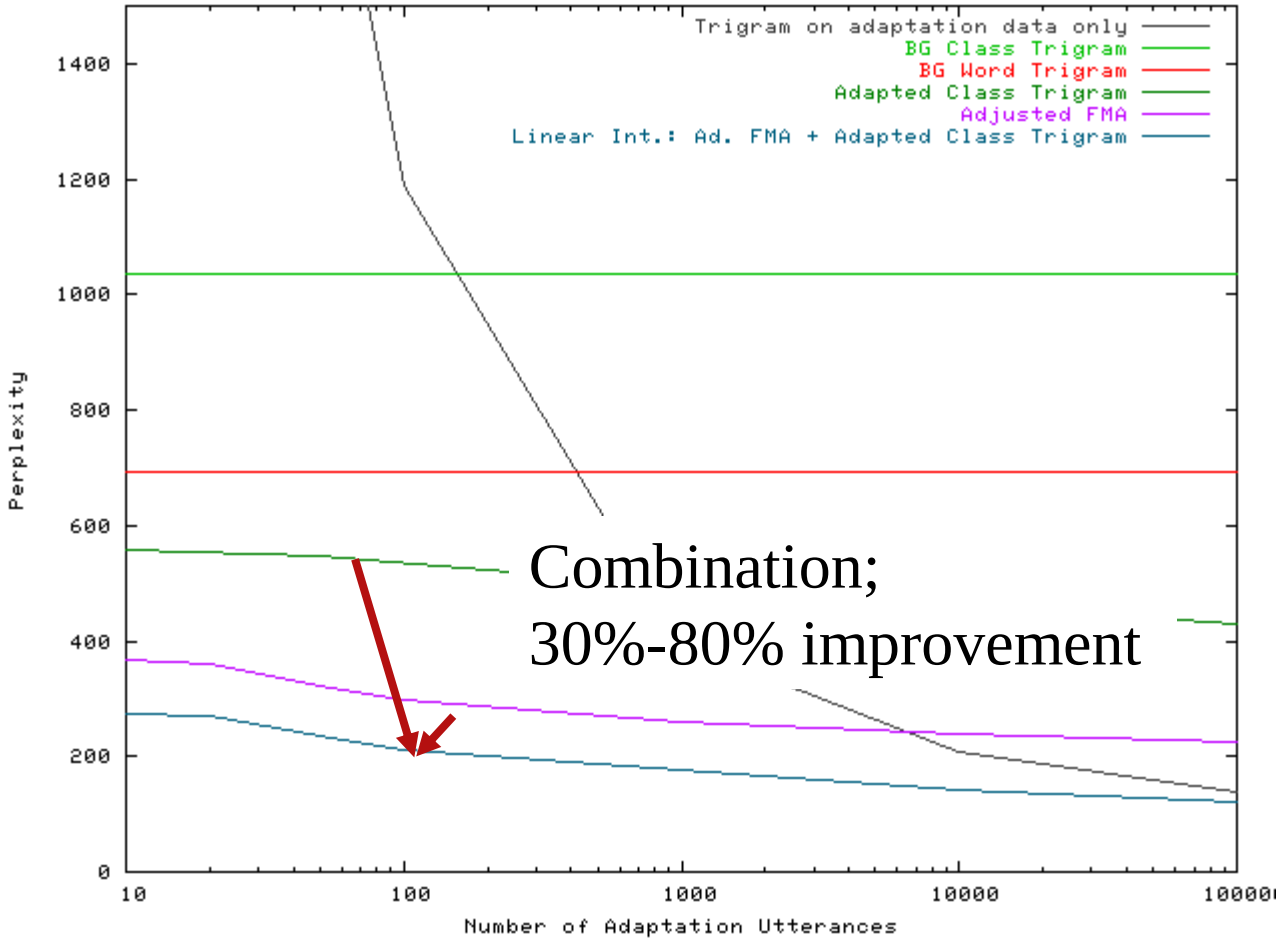
Comparison on Trigram Models



Comparison on Trigram Models



Comparison on Trigram Models



Summary

- Demonstrated improvements on
  - Fill-up
  - FMA
- Suggest adaptation of class LMs
- Combination of method makes LM adaptation to tiny corpora feasible

Overall Summary

- “Inventing” new words
- Selecting documents to decrease mismatch between background corpus and target domain
- Adapting LMs using only 10 domain specific utterances