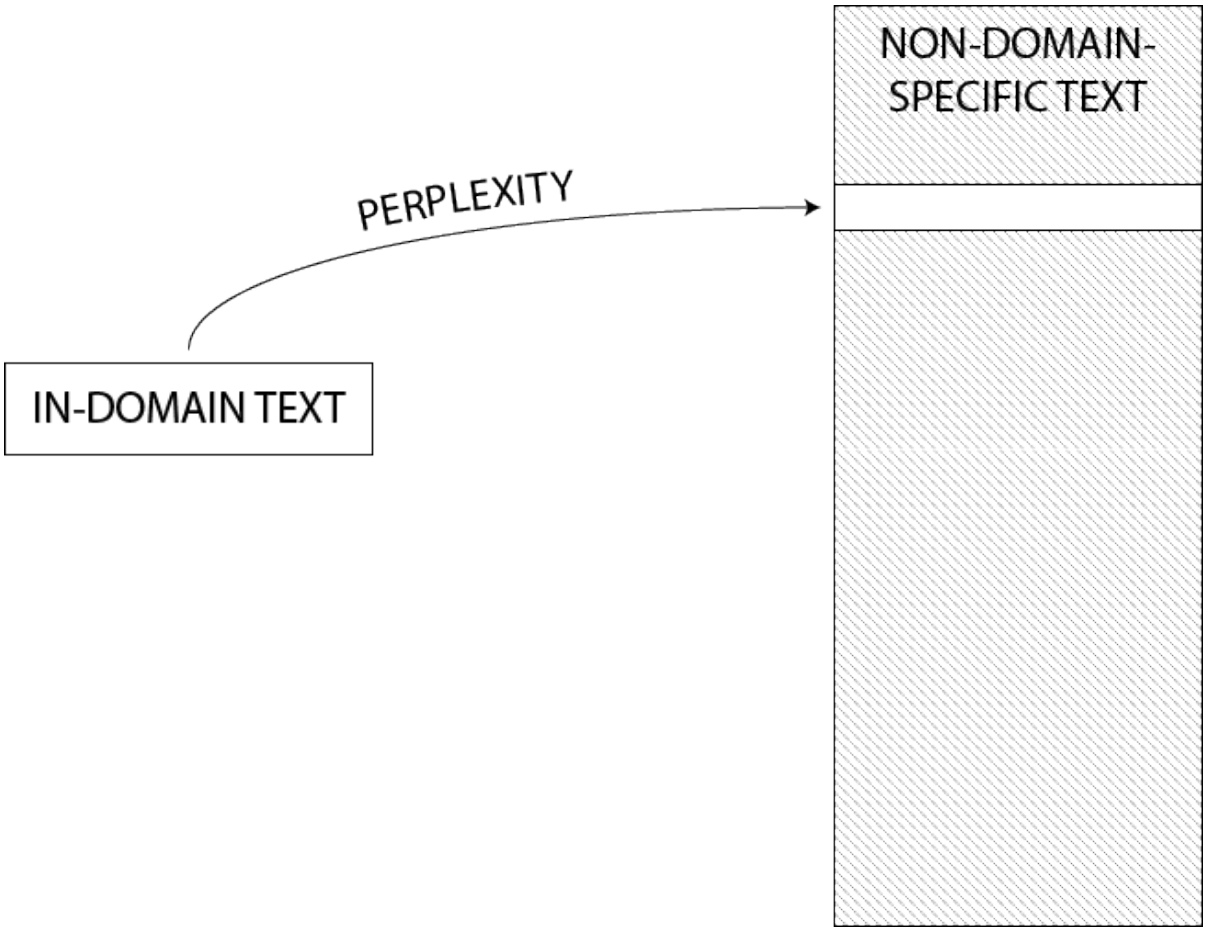
Studies on Training Text Selection for Conversational Finnish Language Modeling

Seppo Enarvi
Aalto University
Espoo, Finland

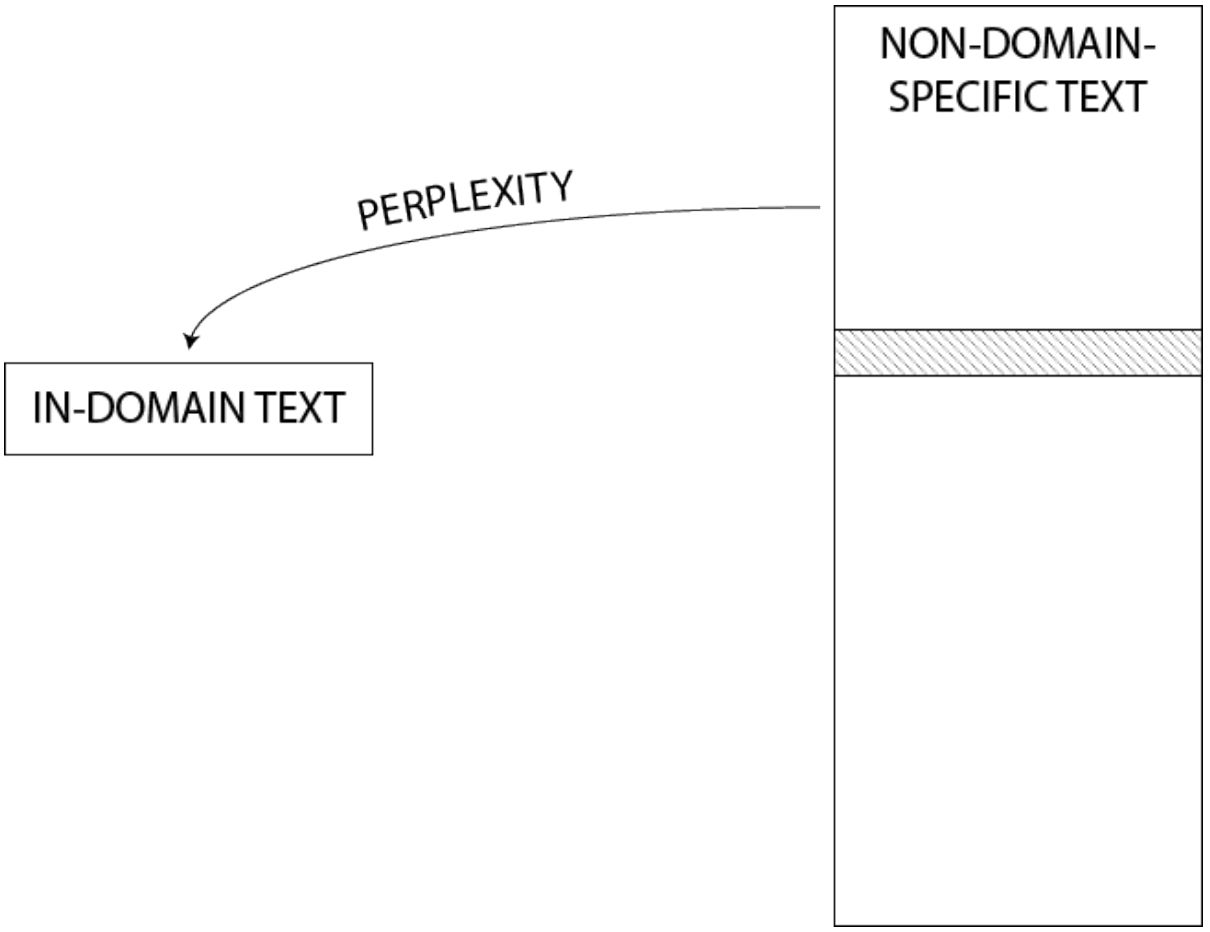
Collected Finnish Conversations

Data set	Number of words
WEB1	767,669
WEB2	1,067,993
WEB3	562,426
WEB4	25,131,015
WEB5	46,258,268
WEB6	2,618,084,259
DEVEL1	17,209
DEVEL2	8,755
EVAL	6,271

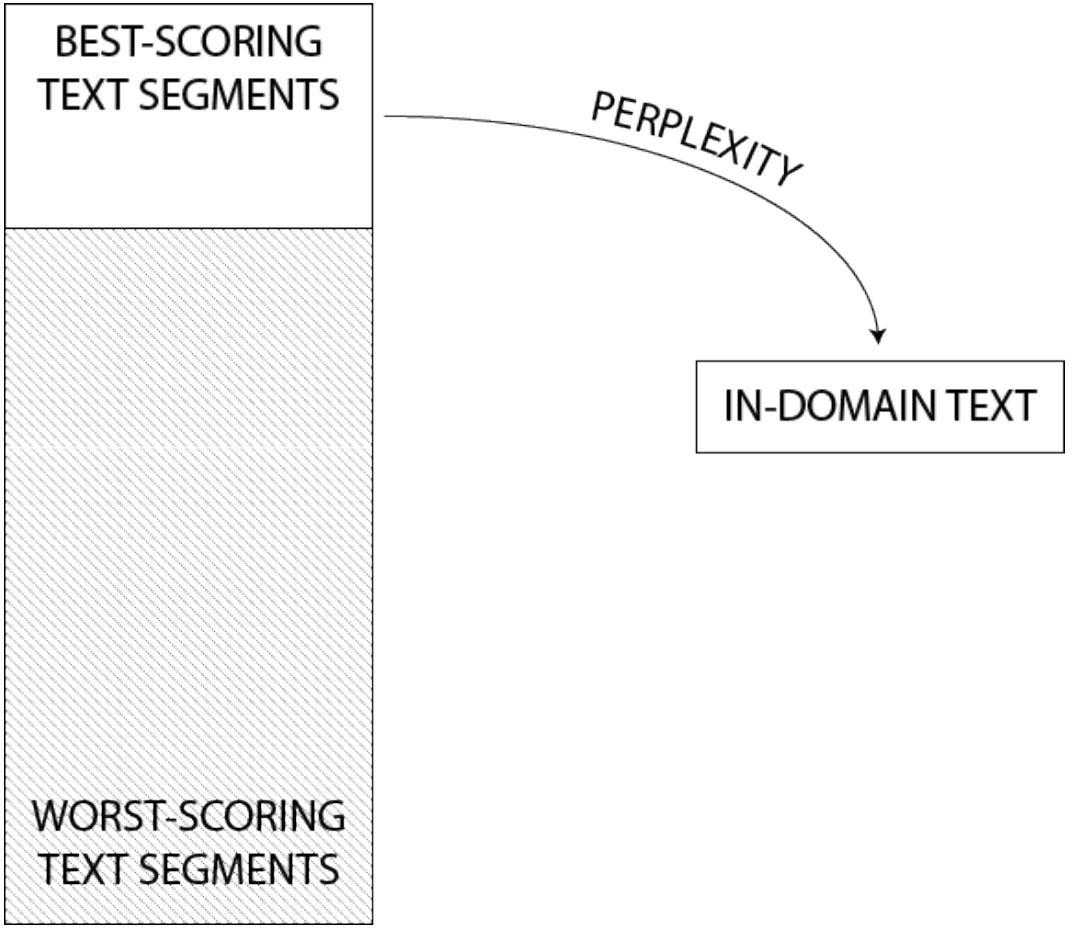
In-Domain LM Perplexity on Training Text



Training Text Perplexity on In-Domain Text



Finding Optimal Filtering Threshold



Phonetic Variation in Finnish Conversations

- Finnish is highly agglutinative.
- Finnish orthography is very close to phonemic.
- In informal conversations, phonetic variation is preserved in writing.

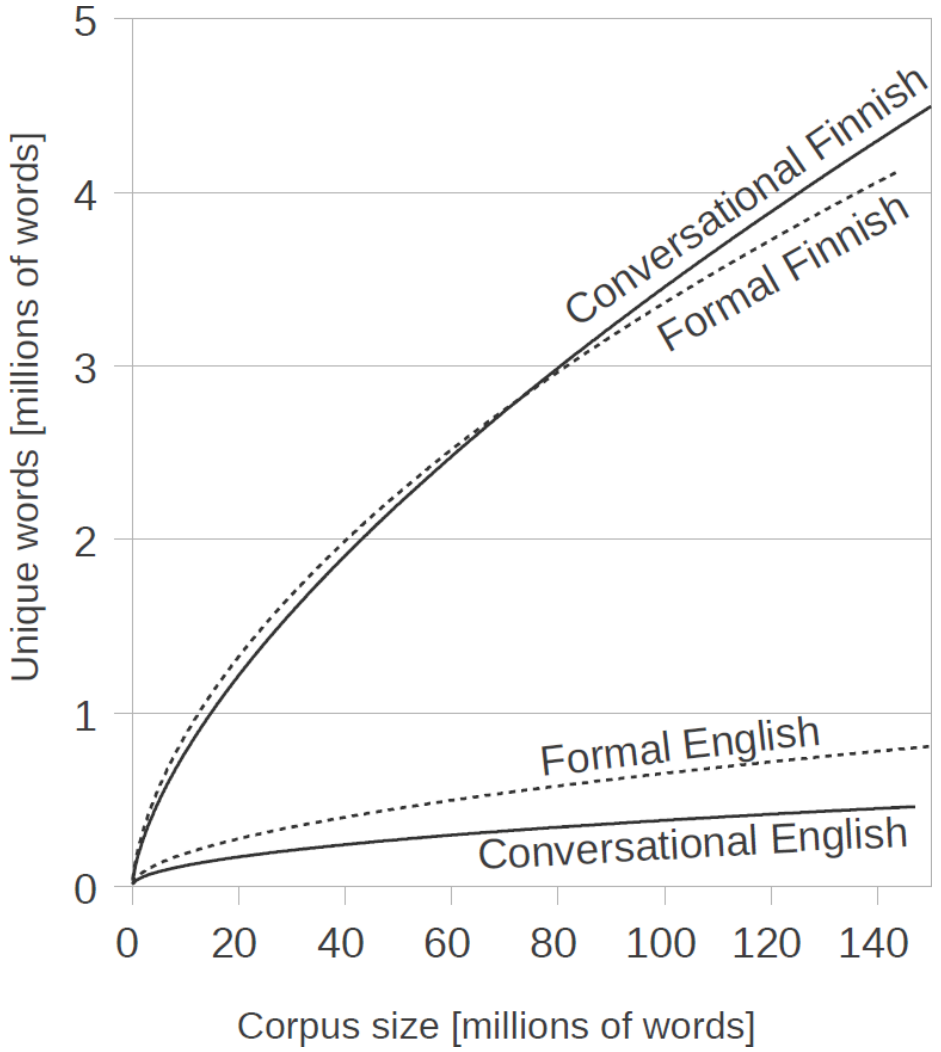
  en minä tiedä

  en mä tiedä

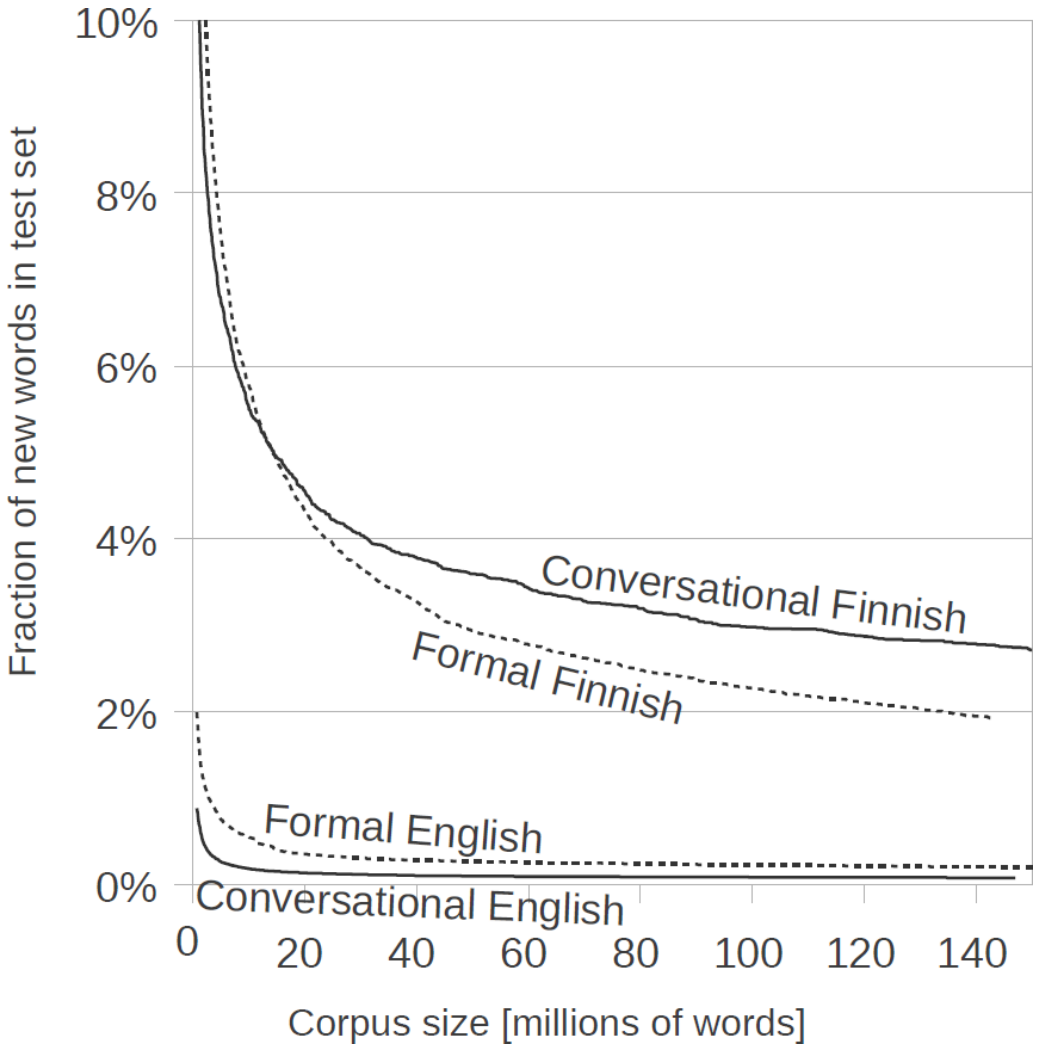
  en mä tiiä

  emmä tiiä

Vocabulary Growth When Corpus Size Increases



Development of OOV Rate When Corpus Size Increases



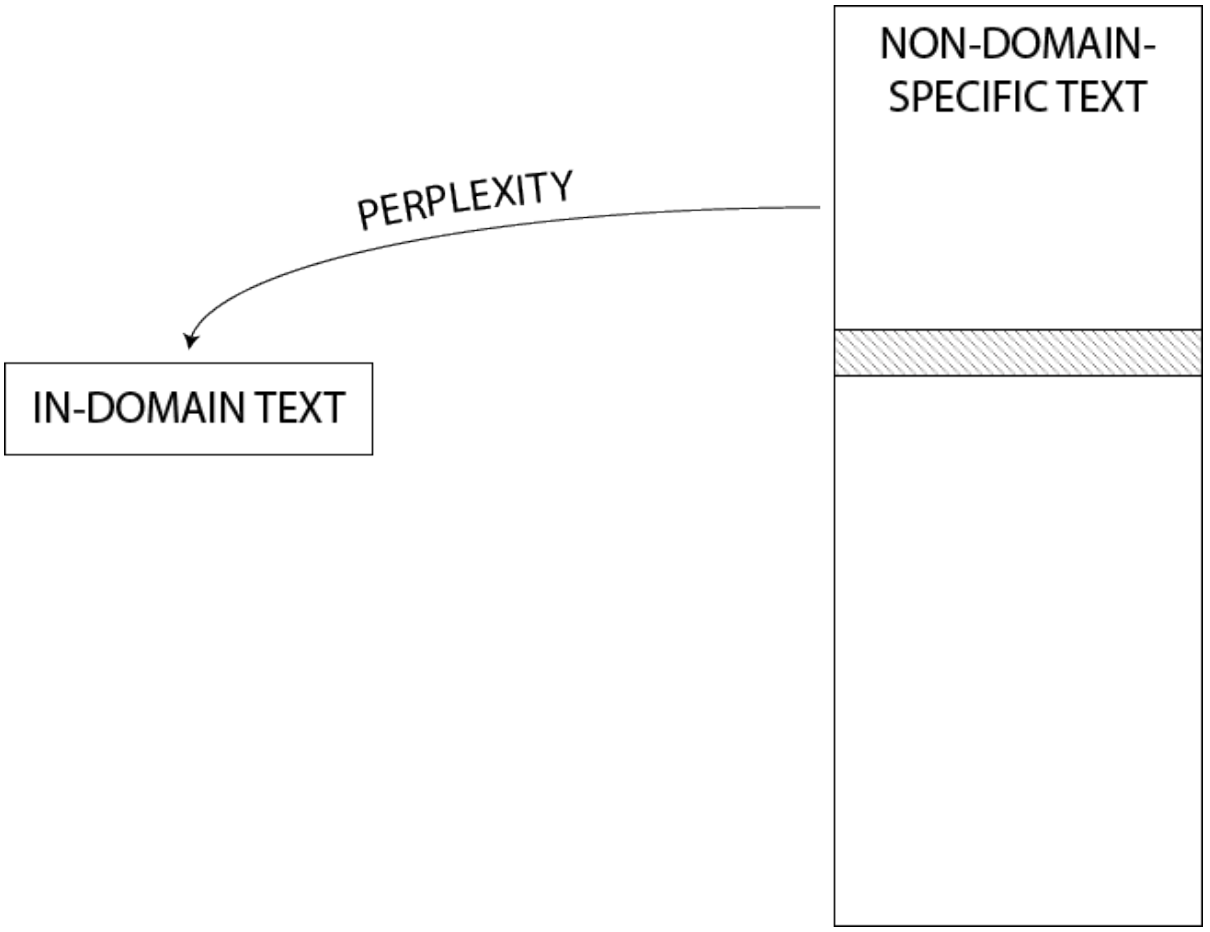
OOV Words in Perplexity Computation

- Ignore OOV words:
  - Higher number of OOV words means lower perplexity value.
- Open-vocabulary language model:
  - Need a reliable estimate for <*UNK*> token probability.
  - A lot of words forms occur only once. This is not a good choice!

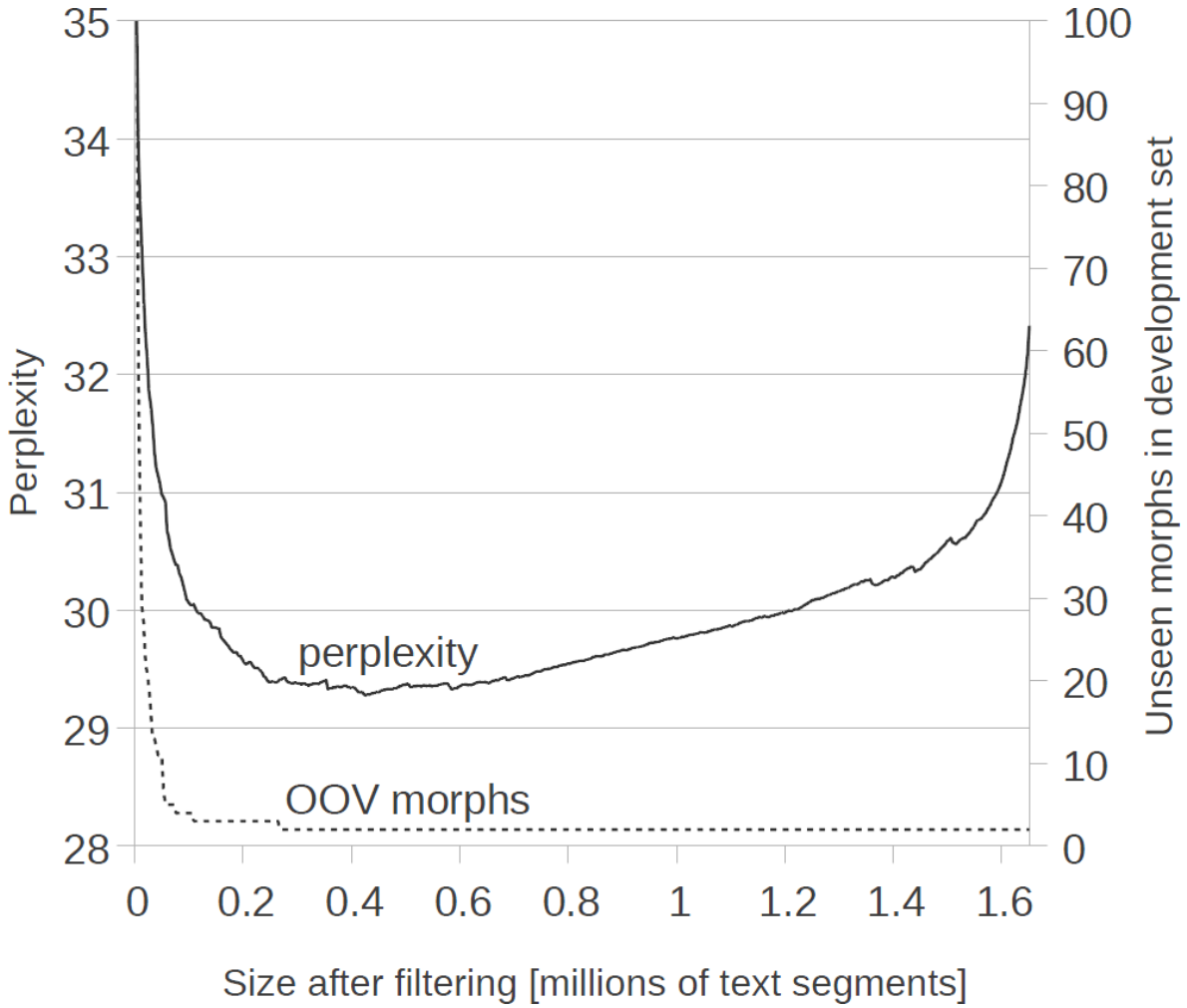
Morph-Based Perplexity Computation

- Unfeasible to re-segment the data at every perplexity computation.
- If only a small portion of the data is used to train a language model, OOV morphs may be a problem in perplexity computation.
- OOV rate will be small if the morph segmentation model is learned using only the language model training data (and used to segment the rest of the data).

Text Segment Scoring



Filtering Threshold Optimization



Experiments

- Aalto ASR system
- Acoustic model trained on planned speech from SPEECON + 3 hours of conversational data
- 88,000 word vocabulary
- 44 minutes of evaluation data from 17 speakers

Results

Training data	N-grams	WER	Perplexity
FTC	20,780,423	72.2	6364
CSC	8,772,995	59.8	674
WEB	15,803,759	59.2	652
WEBfilt	3,694,060	57.5	589
CSC+WEB	14,884,046	55.6	493
CSC+WEBfilt	5,429,240	55.7	496

FTC = Finnish Text Collection (Literary Finnish)

CSC = Three corpora available from CSC—IT Center for Science in Finland

WEB = Collected web data sets 1–5

WEBfilt = Filtered web data sets 1–5

Morph-Based Recognition Results

Training set	N-grams	WER
CSC	8,772,995	63.9
CSC+WEB	14,884,046	58.8
CSC+WEBfilt	5,429,240	59.4

CSC = Three corpora available from CSC—IT Center for Science in Finland

WEB = Collected web data sets 1–5

WEBfilt = Filtered web data sets 1–5

Future Work

- Filtering of another 23 GB of Internet conversations.
- Find out why morph-based models fail in conversational Finnish ASR.
- Use the filtered data to find clusterings of word forms.

Thank You!

- Code has been released in GitHub:

  https://github.com/senarvi/senarvi-speech/
- New corpora available on request.
- E-mail: seppo.enarvi@aalto.fi