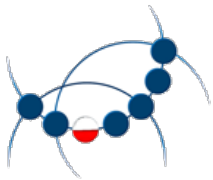
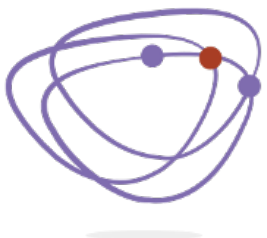
Preparing a speech corpus using the recordings of the Polish Film Chronicle

Danijel Koržinek

Polsko-Japońska Akademia Technik Komputerowych





CENTRUM TECHNOLOGII
JĘZYKOWYCH CLARIN-PL

11 June 2018, Warsaw

Introduction

About Clarin-PL

- K-type Clarin centre operating in Poland since 2013
- Run by teams from 6 Polish universities:
  - Wrocław University of Science and Technology
  - Institute of Computer Science PAS
  - PJAIT
  - Institute of Slavic Studies PAS
  - University of Łódź
  - Wrocław University
- It deals with various topics, including:
  - computer linguistics, social linguistics, language translation, language history and speech
- http://clarin-pl.eu

PJAIT

- Polish-Japanese Academy of Information Technology in Warsaw
- Tasks within the Clarin-PL project include:
  - creation of speech resources
  - development of tools and services for automatic speech processing
  - long term archive

Motivation

- The Chronopress toolkit
- Other speech corpora (eg. parliamentary speech)
- Public domain?
- Type of data:
  - narrow domain
  - (mostly) single speaker
  - “predictable” audio
- Usefulness as a resource

Polish Film Chronicle

- Examples PKF vs British
- Advantages:
  - mostly (ie. 90%) single speaker
  - narrow and fairly well defined domain
  - uniform quality
- Disadvantages:
  - low quality, noise, disruptions
  - high level of background music
  - lots of “named entities”

Aim and scope

- To create a corpus of PKF spanning the years 1945-1962
- Only narrator's speech is required
- Create a speech recognition system that is capable of recognizing this type of audio

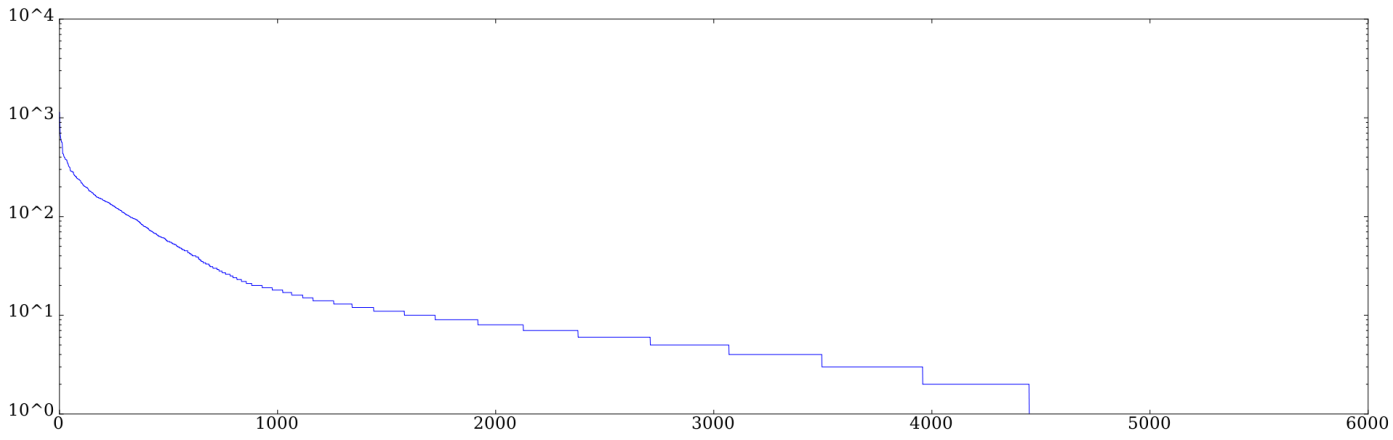
Data acquisition

First data source

- KronikaRP.pl - now defunct
  - Videos (in MP4 format)
  - Transcriptions (only the narrator)
  - Metadata (production team, location, date, etc...)
- Custom crawler (Python + Beautiful Soup)

Data prepossessing

- Downloaded 5585 files
- Mean length 97s
- Filtered out bad sound quality and transcription leaving 5123 files
- Used standard GMM model to align - 1815 files
- Retrained and realigned - 5098 files
- Retrained and `find_bad_utts.sh`



Error summary

- Out of 5123 files
- 283 have no errors
- 677 have <= 1 error
- 1165 have <= 2 errors
- 1627 have <= 3 errors
- 2053 have <= 4 errors
- 3541 have <= 10 errors
- 4242 have <= 20 errors
- Worse 10 have these errors: 1138, 1006, 846, 751, 660, 652, 618, 594, 588, 569

WER sample

```
AndLap_10958 ref  *** na znak protestu przeciwko
  ↪ powrotowi na tron króla kolaboracjonisty
  ↪ leopolda trzeciej *** setki tysięcy belgów
  ↪ demonstrują na ulicach brukseli ...
AndLap_10958 hyp  w na znak protestu przeciwko
  ↪ powrotowi na tron króla kolaboracjonisty
  ↪ leopolda trzeciej po setki tysięcy belgów
  ↪ demonstrują na ulicach brukseli ...
AndLap_10958 op   I   C   C       C         C
  ↪ C       C   C     C           C           C
  ↪ C     I     C       C       C       C       C
  ↪ C         C   ...
AndLap_10958 #csid 65 1 5 0
```

Manual correction

- Chose 400 worst files
- Manually corrected by 4 people
- Realigned and resorted
- Picked 400 worst again and fixed manually

FN corpus

- Collection of 4373
- Mean length 73s
- Loosely related to KronikiRP data
  - Out of 560 chronicles, 195 only in FN and 79 only in KronikaRP
- Order within chronicles not guaranteed
- Used decoding trick to match KronikaRP transcriptions to FN
- 2796 transcriptions matched - 1587 lacked transcription
  - They were manually corrected from ASR output

Theoretical background

Automatic Speech Recognition

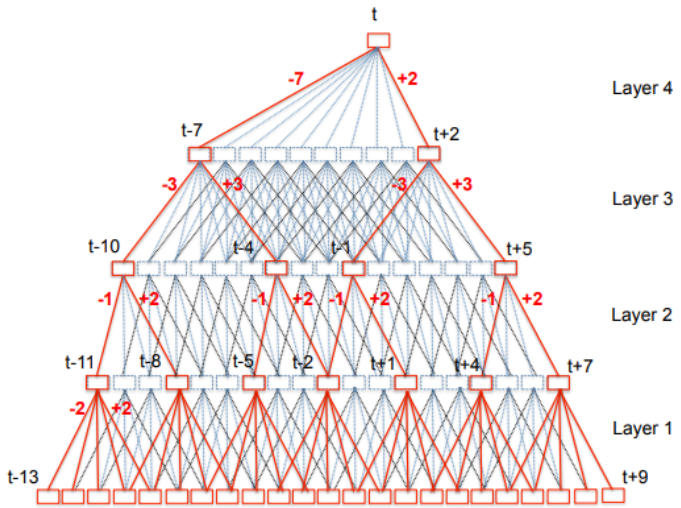
- The goal is to find the optimal word sequence using the Bayes rule:

$$\arg\max_{w} P(w|O) = \arg\max_{w} P(O|w) \cdot P(w) \tag{1}$$

- Acoustic model – $P(O|w)$
  - GMM+HMM, ANN+HMM
  - GMM+WFST, DNN+WFST
  - end-to-end, CTC
- speaker adaptation, VAD
- Language model – $P(w)$
  - FSA, graph models
  - n-gram, statistical models
  - ANN-based, RNNLM, LSTMLM, ...
- 1-best, n-best, lattices, rescoring
- We use Kaldi speech recognition toolkit for all our experiments

Acoustic model

- Standard recipe:
  - train mono, realign, (train tri, realign)x2, LDA+MLLT, SAT (fMLLR)
  - TDNN, (B)LSTM, chain models
- Chain models:
  - output framerate = 1/3 input framerate
  - sequence-level objective function (MMI on phone n-gram decoding result)
- adaptation and transfer learning



VAD

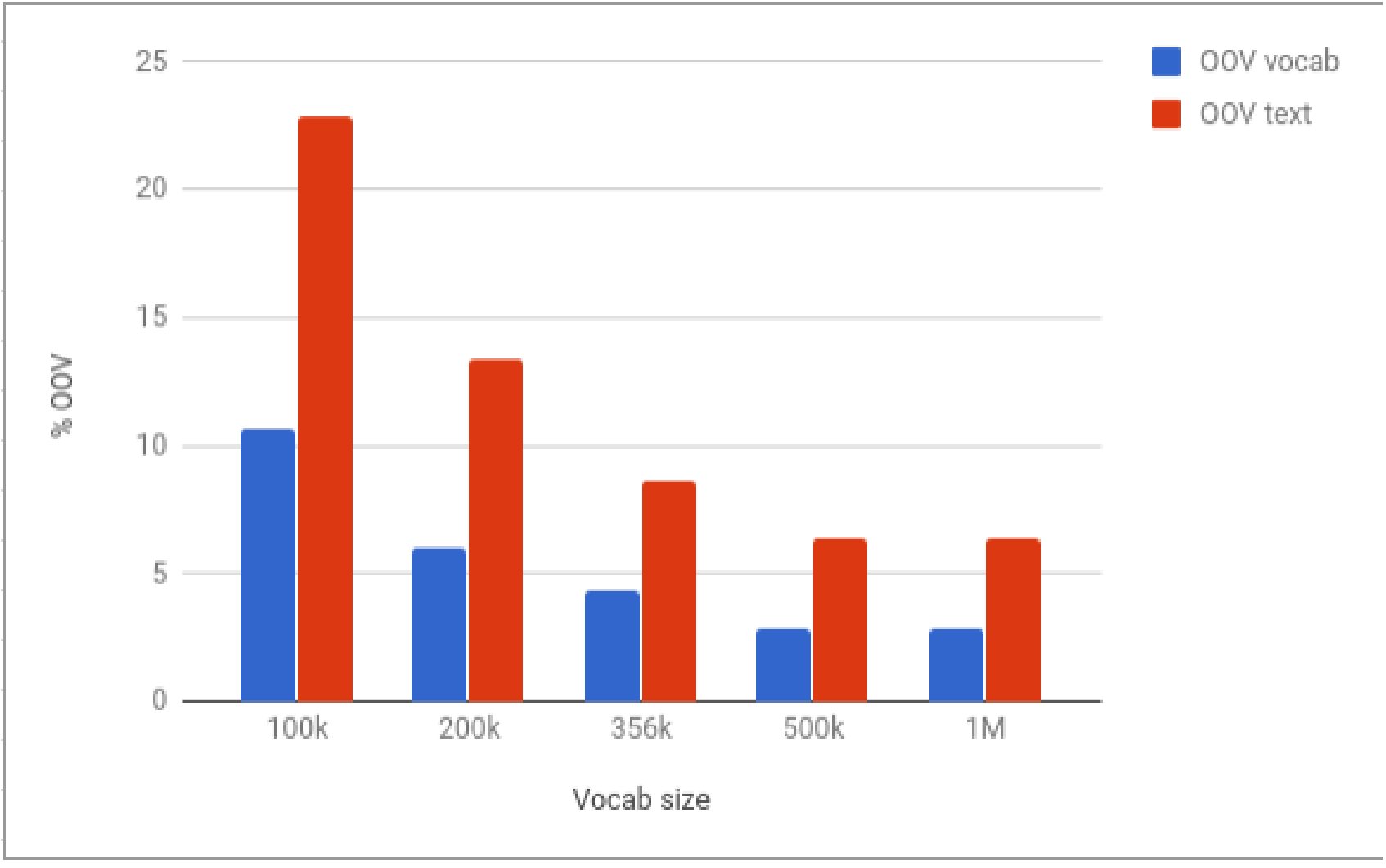
- binary classification problem on individual (independent) frames of audio
- feed-forward network with 5-frame input (center + 2 previous + 2 next)
- 64 mel-filters in input
- 3 hidden layers: $1024 \rightarrow 512 \rightarrow 256$
- RELU activations, dropout between hidden layers
- MSE loss and SGD optimizer
- smooth output (10 frame median filter) and 0.5s margins around speech segments
- reached 94% accuracy on development data

Language models

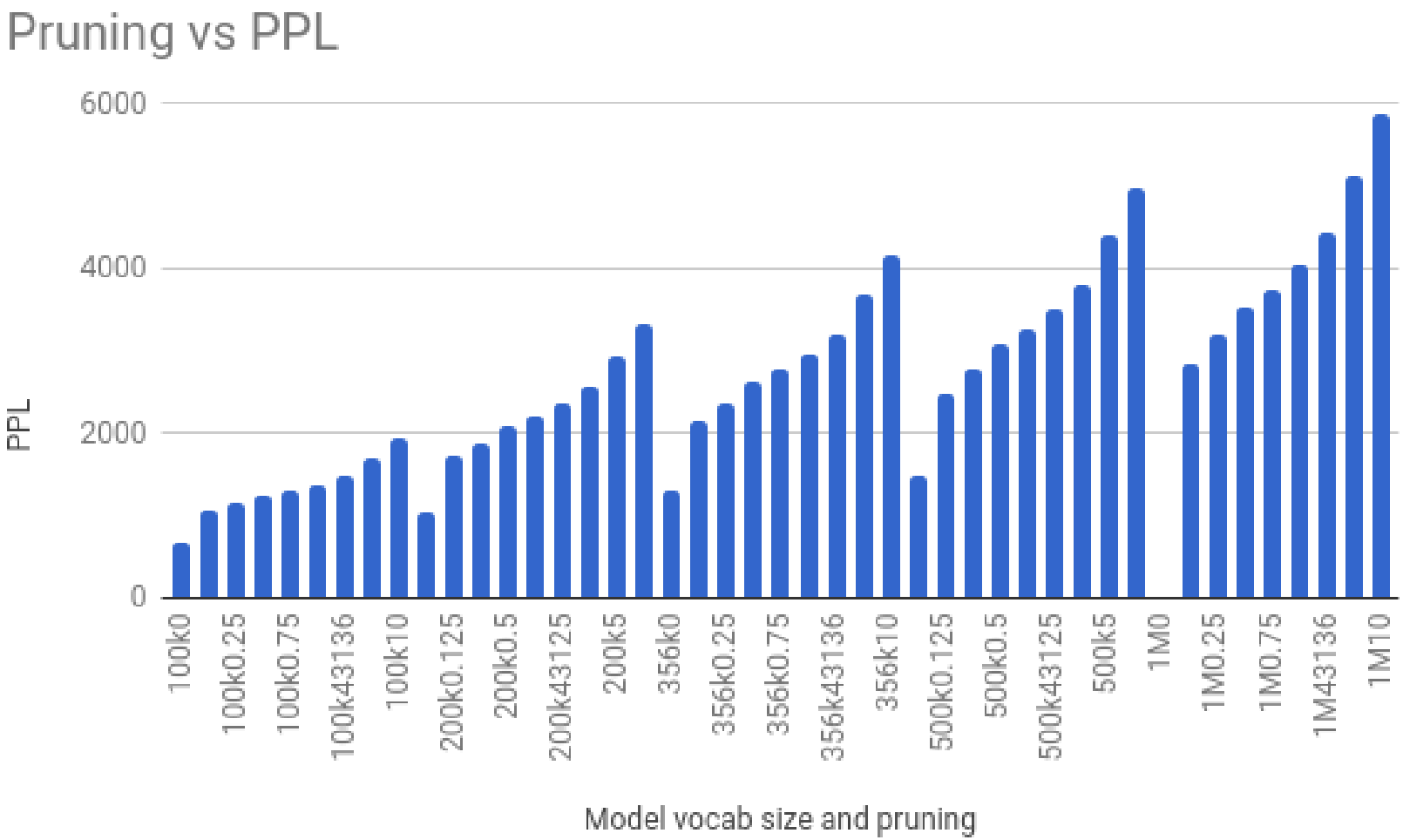
- 3-gram models, Knesser-Ney smoothing
- initally SRILM, but finally PocoLM:
  - https://github.com/danpovey/pocolm
  - optimizes interpolation and hyperparametr tuning
  - provides an optimal LM and vocabulary to fit development data
- lattice rescoring
- neural models

Experiments

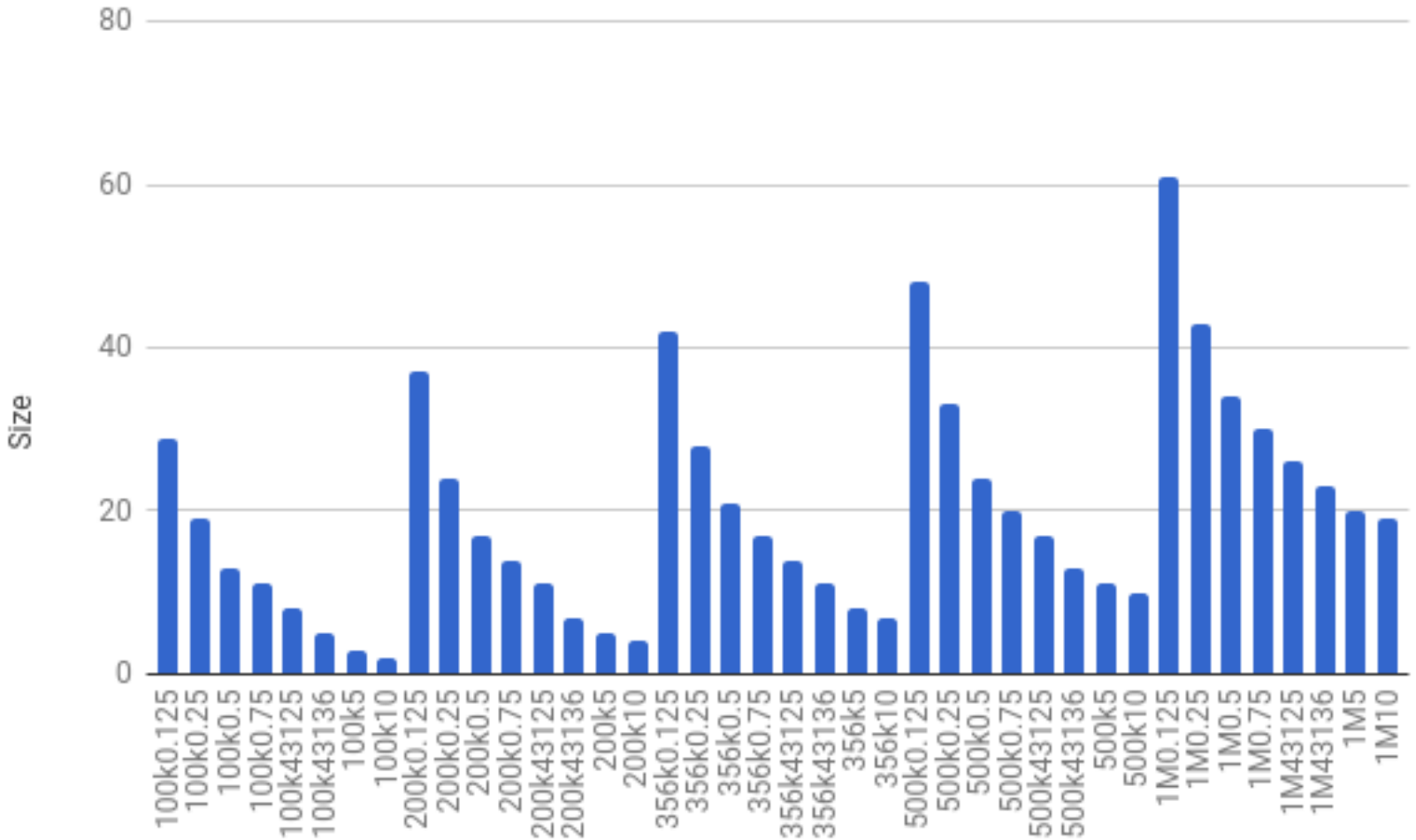
Vocabulary and OOV



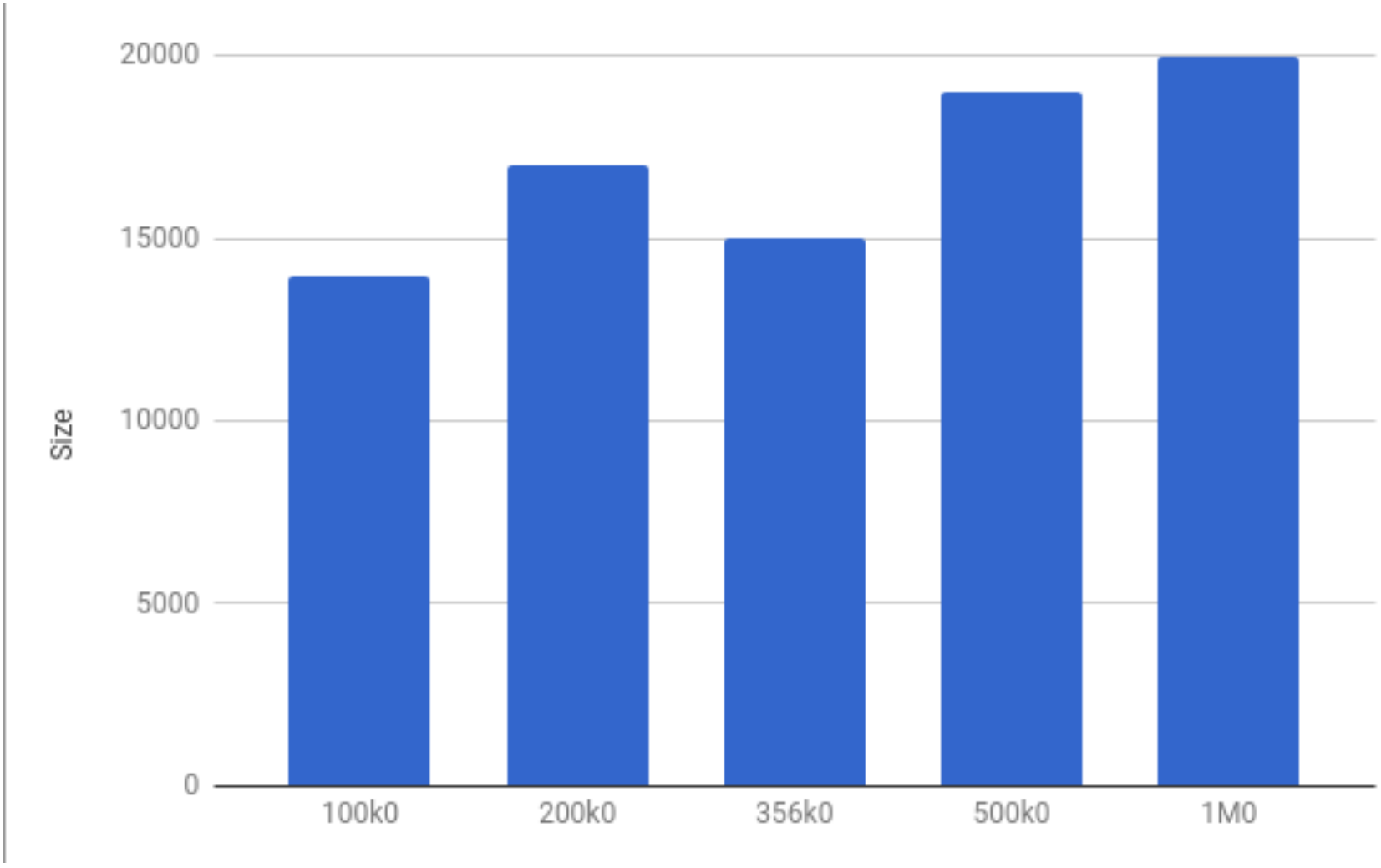
Pruning vs PPL



Pruning vs Size



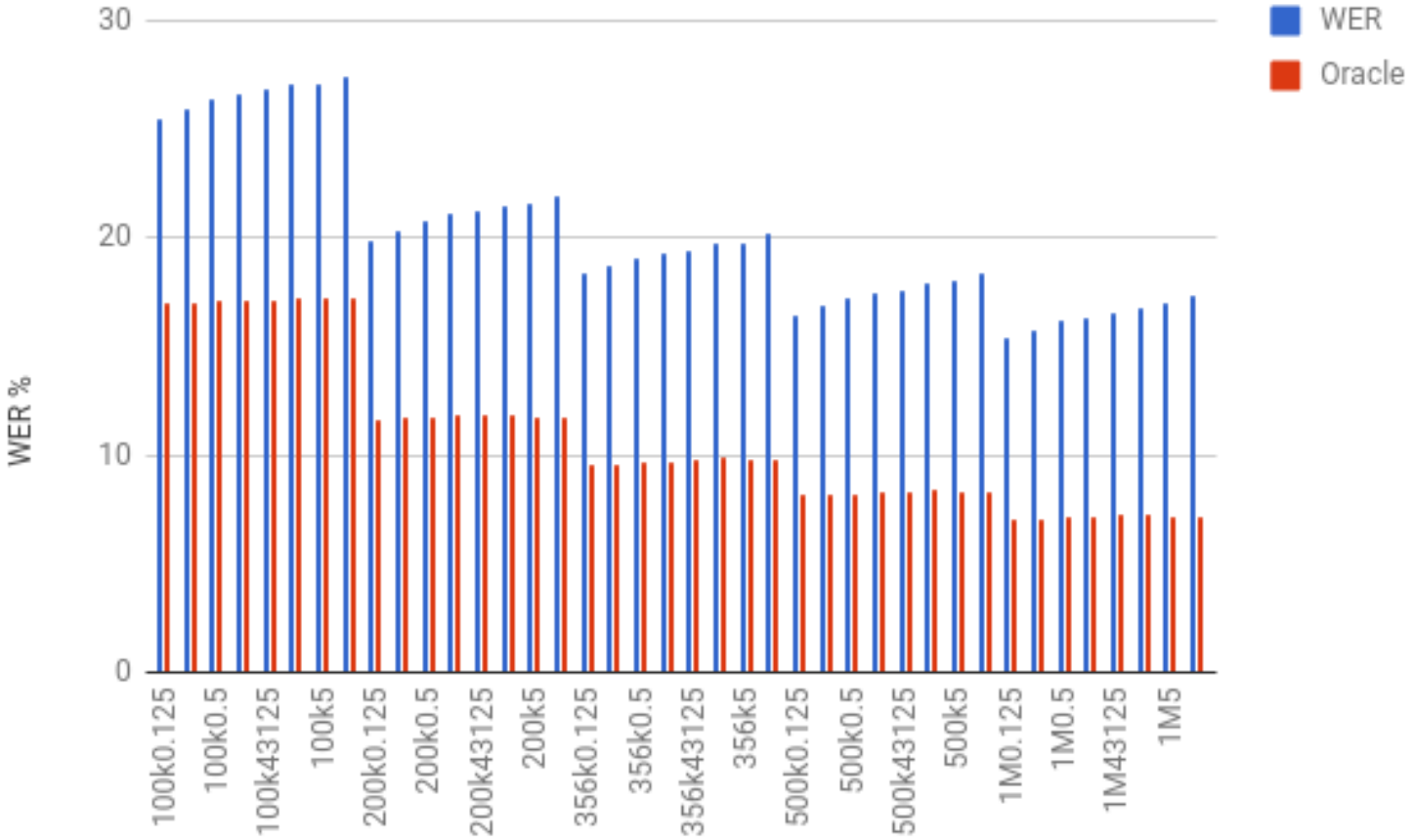
Size with no pruning



HCLG.fst

- HCLG:
  - H – HMM-level states
  - C – state context
  - L – lexicon, ie. phoneme to word conversion
  - G – graph, ie. language model
- L - proportional to lexicon size (4k words per MB)
- G - proportional to LM size (approx. 1:2)
- HCLG - mostly dependent on G (approx 10:1)
- HCLG were not created for LMs with no pruning!

Pruning vs WER

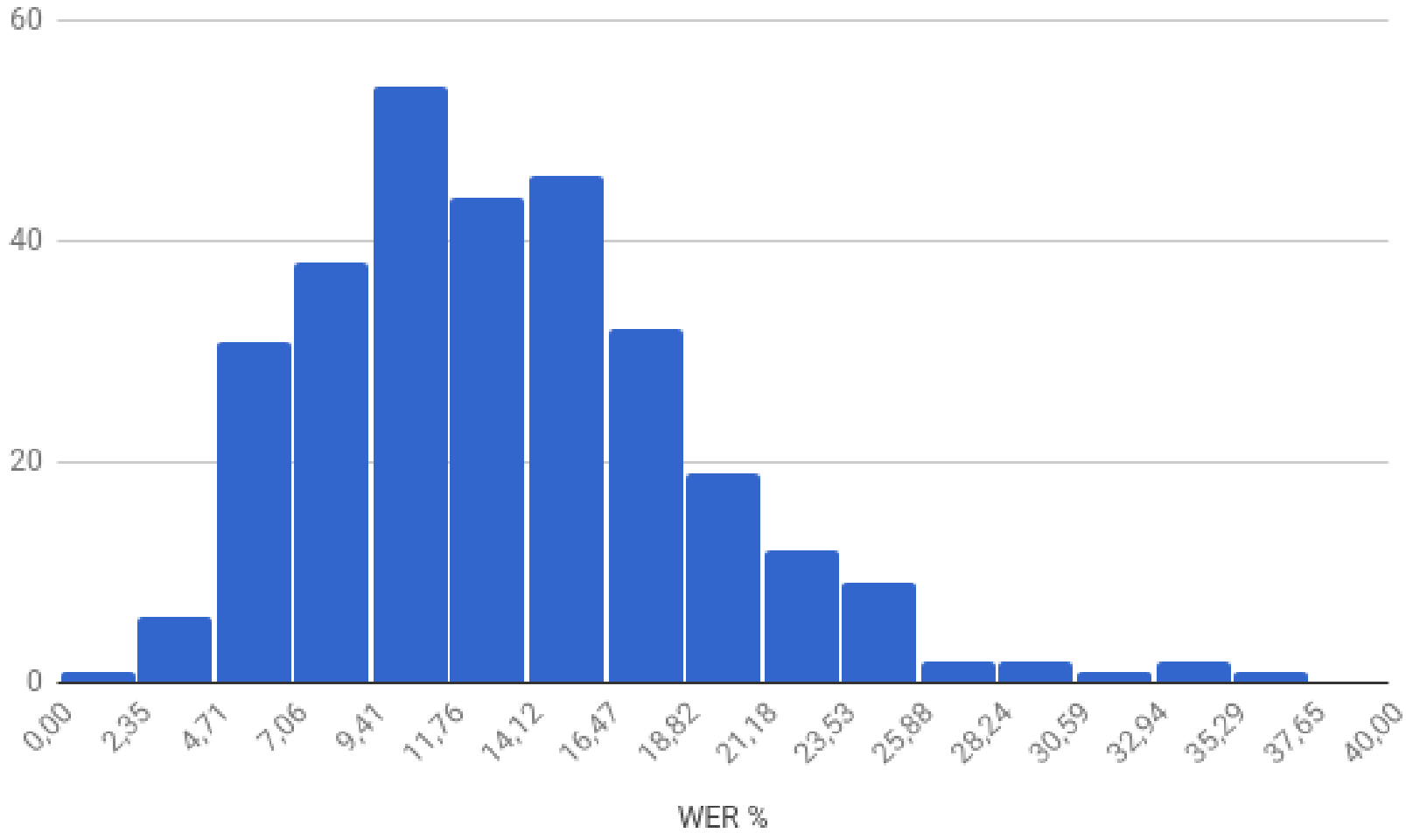


- Train transcritpions WER: 27.27% Oracle: 18.79%
- Train+test transcritpions WER: 5.67% Oracle: 3.67%

Lattices and rescoring

- We took largest vocab (1M) and modest pruning (0.125)
  - Regular – WER: 15.35% Oracle: 7.05%
  - Double beam (30/16) – WER: 15.32% Oracle: 6.26%
- Rescored using biggest model (1M, no pruning, 20G size)
  - Regular rescored – WER: 13.54%
  - Wide beam rescored – WER: 13.58%
- For comparison
  - GMM-based model using same LM – WER: 33.93% Oracle: 20.61%
  - Non-adapted chain model – WER: 68.16% Oracle: 45.85%
  - Google Cloud Speech – WER: 17.96%

WER histogram



(wide beam rescored model)

VAD experiments

- We expect that removing music will reduce WER
- No VAD:
  - Double beam (30/16) – WER: 15.32% Oracle: 6.26%
  - Wide beam rescored – WER: 13.58%
- VAD:
  - Using oracle VAD – WER: 12.34% Oracle: 3.33%
  - Not using margins – WER: 19.02% Oracle: 5.28%
  - With margins – WER: 12.67% Oracle: 3.46%
  - Rescored – WER: 10.97%

Summary

- Non-adapted AM, adapted LM – WER: 68.16%
- GMM-based AM, adapted LM – WER: 33.93%
- Adapted AM, weak LM – WER: 27.27%
- Adapted small LM – WER: 15.35%
- Wider beam – WER: 15.32%
- VAD – WER: 12.67%
- Rescoring – WER: 10.97%
- Google – WER: 17.96%

Conclusions

- AM adaptation is important
- No reason to use GMM models for transcription
  - Still useful for alignment
- LM adaptation is even more important
- Vocabulary size is not a bottleneck!
- LM size can be a bottleneck
- Lattice rescoring on a wider beam can alleviate that problem
  - Obviously works with offline transcription only
- VAD is often very useful

Questions

danijel@pja.edu.pl