The CSTR System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages

Ondrej Klejch, Electra Wallington, Peter Bell

Centre for Speech Technology Research, University of Edinburgh

Subtask 1 System Overview

Acoustic Model

- CNN-TDNN with language specific outputs
- Trained with LF-MMI
- Multilingual pre-training
- Monolingual fine-tuning

Language Model

- Language specific 3-gram and RNN LMs
- Training data + CommonCrawl data

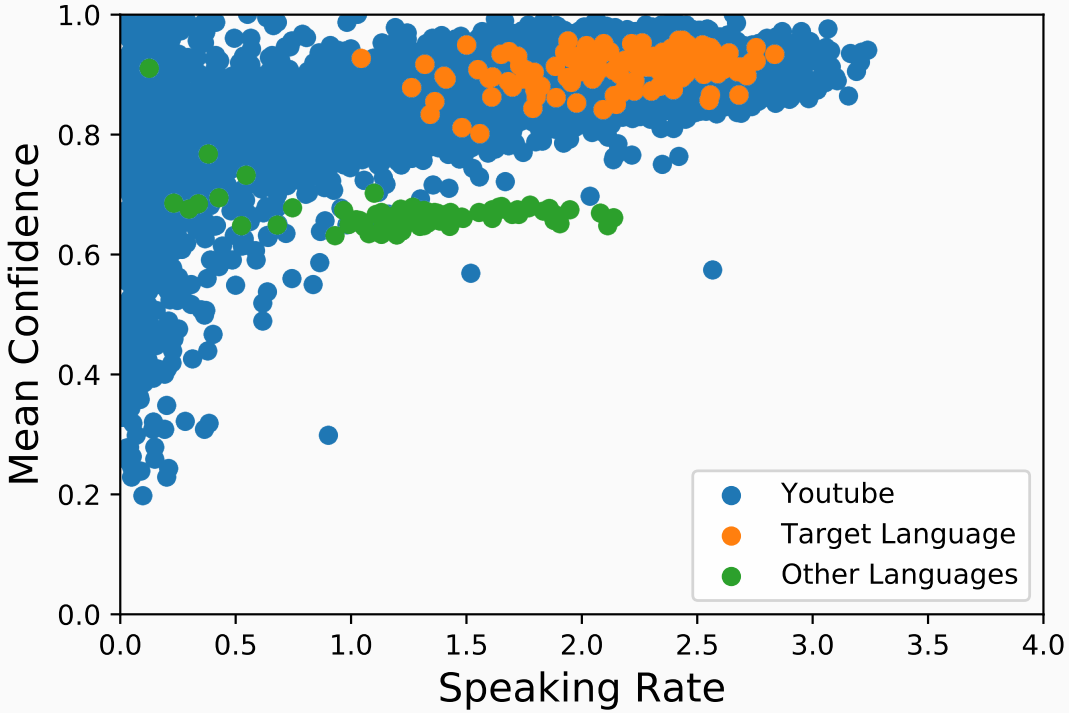
Youtube Crawling

To increase the training data size we crawl Youtube videos by searching for the most common trigrams.

Youtube Crawling

To increase the training data size we crawl Youtube videos by searching for the most common trigrams.

We filter videos using mean confidence and speaking rate.



Semi-Supervised Training

Because we do not have transcriptions for Youtube videos, we use Semi-Supervised Training.

Semi-Supervised Training

Because we do not have transcriptions for Youtube videos, we use Semi-Supervised Training.

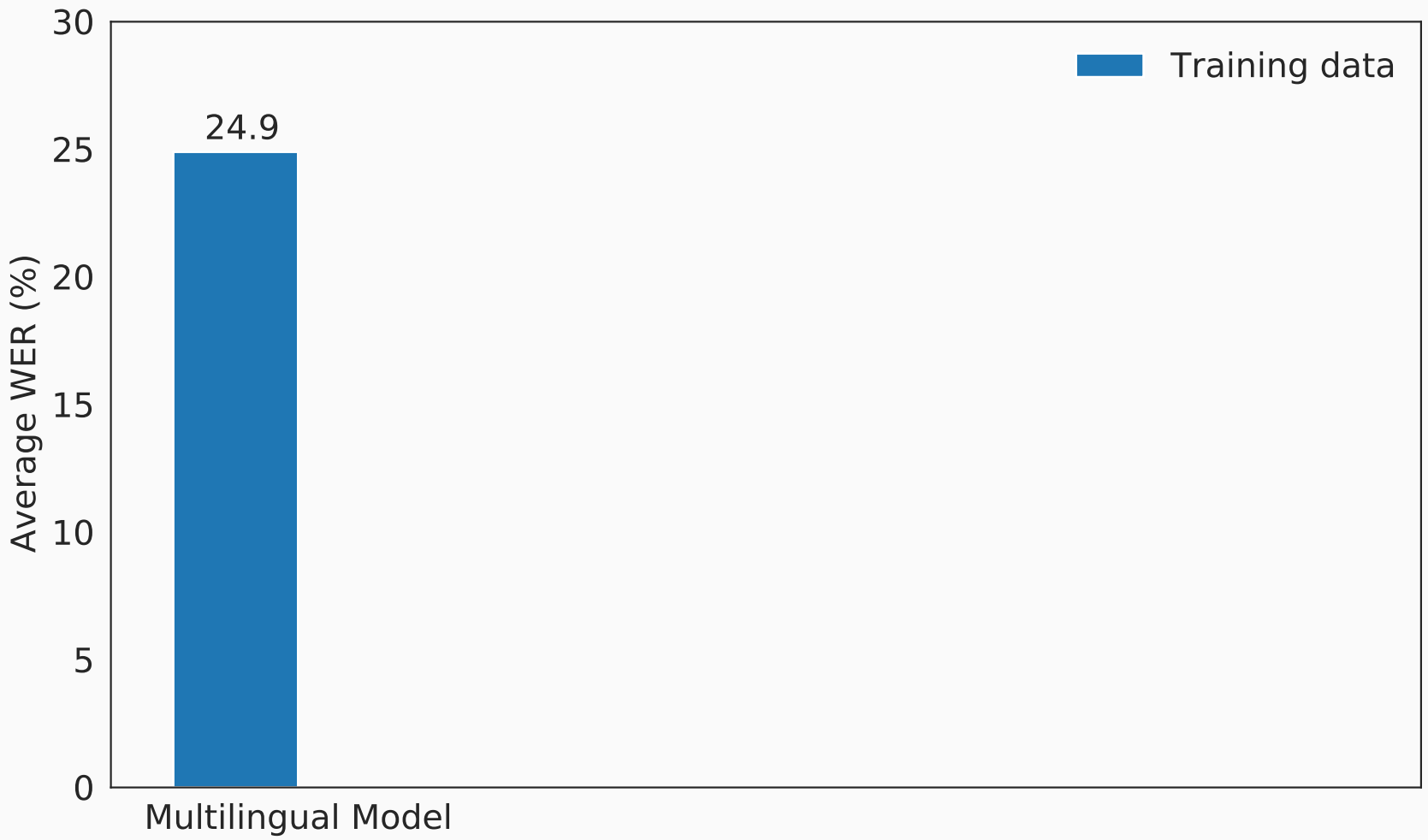
- We decode the videos with a seed model.

Semi-Supervised Training

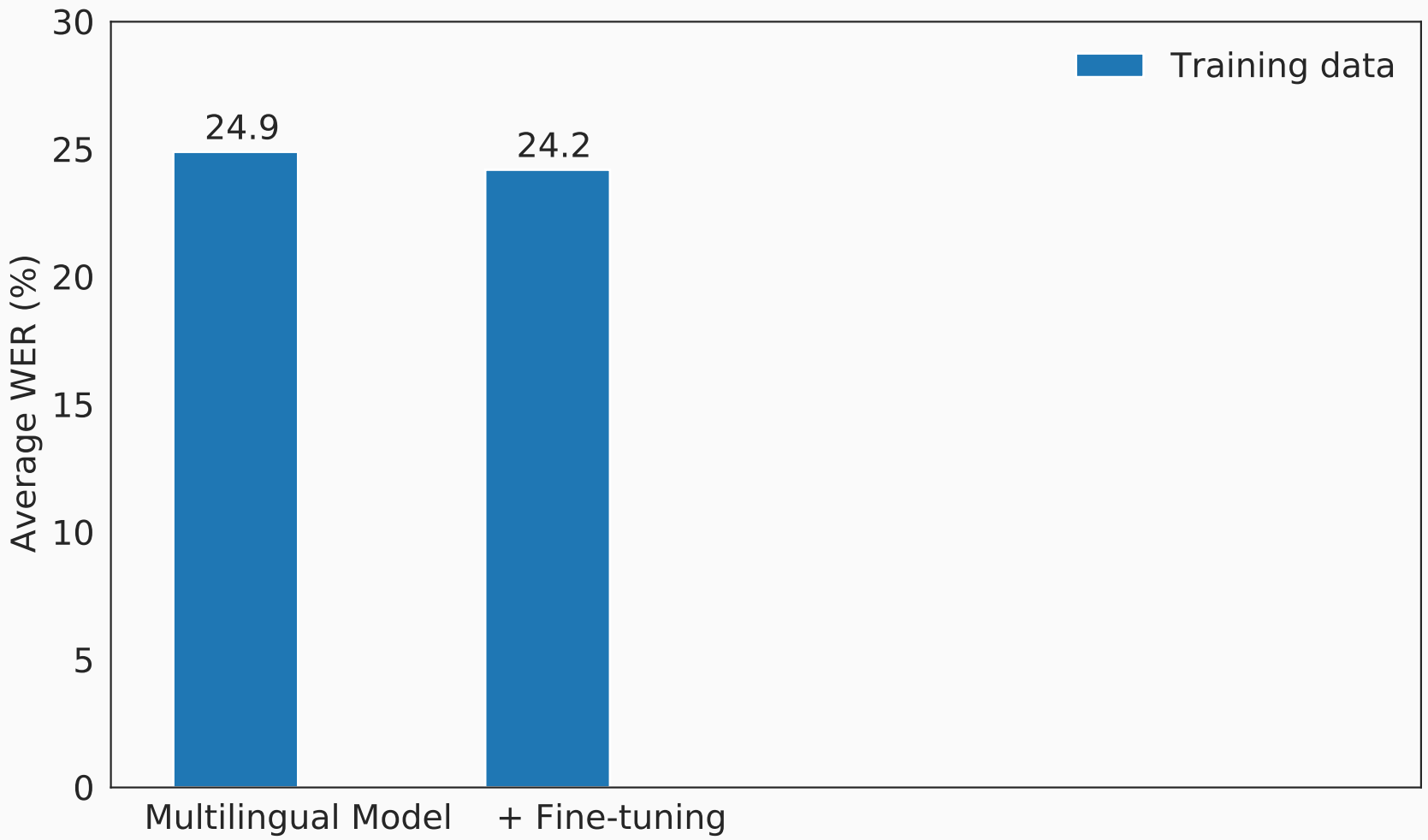
Because we do not have transcriptions for Youtube videos, we use Semi-Supervised Training.

- We decode the videos with a seed model.
- We use the decoded output as labels for training.

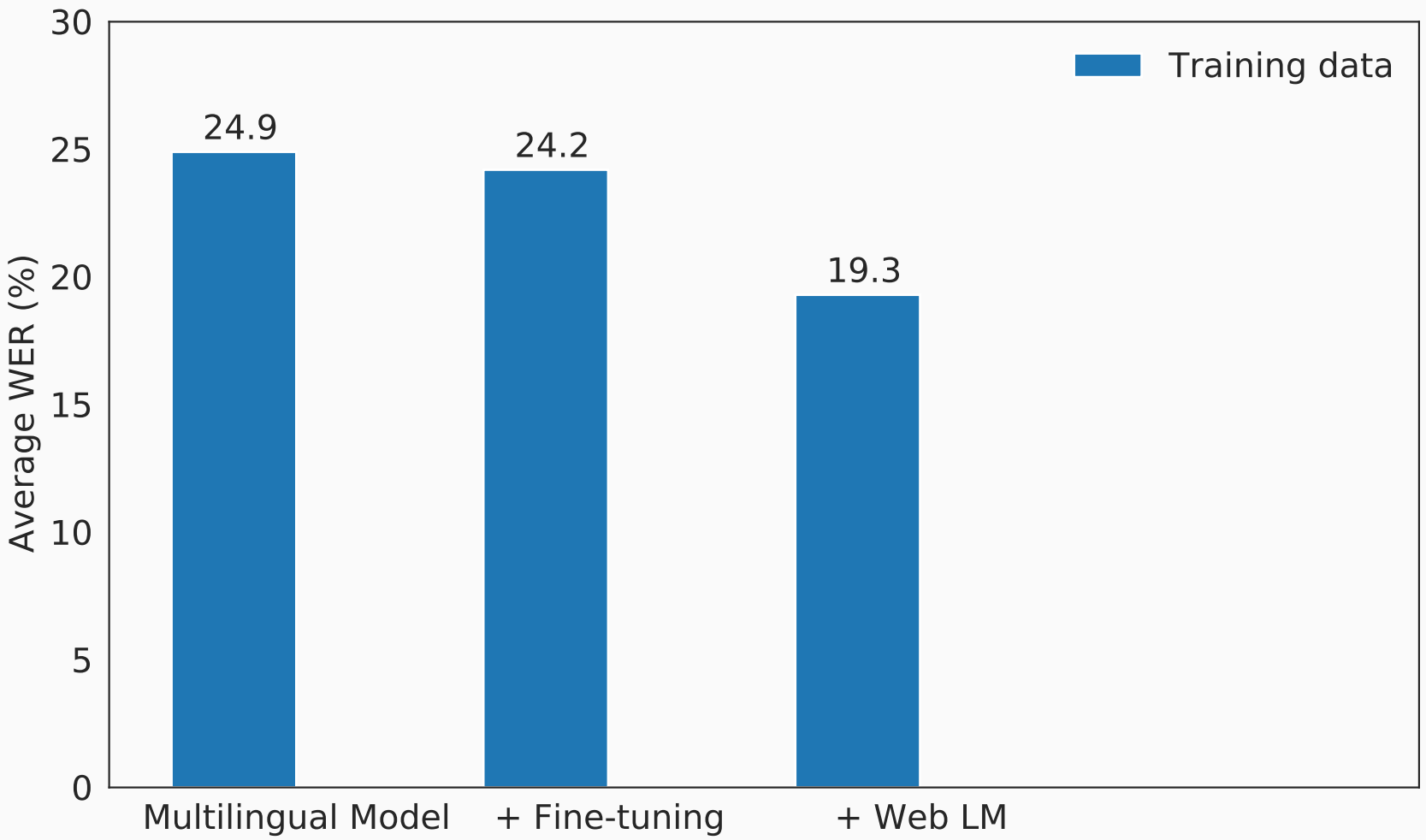
Results


30
25
20
15
10
5
0
Average WER (%)
24.9
Training data
Multilingual Model

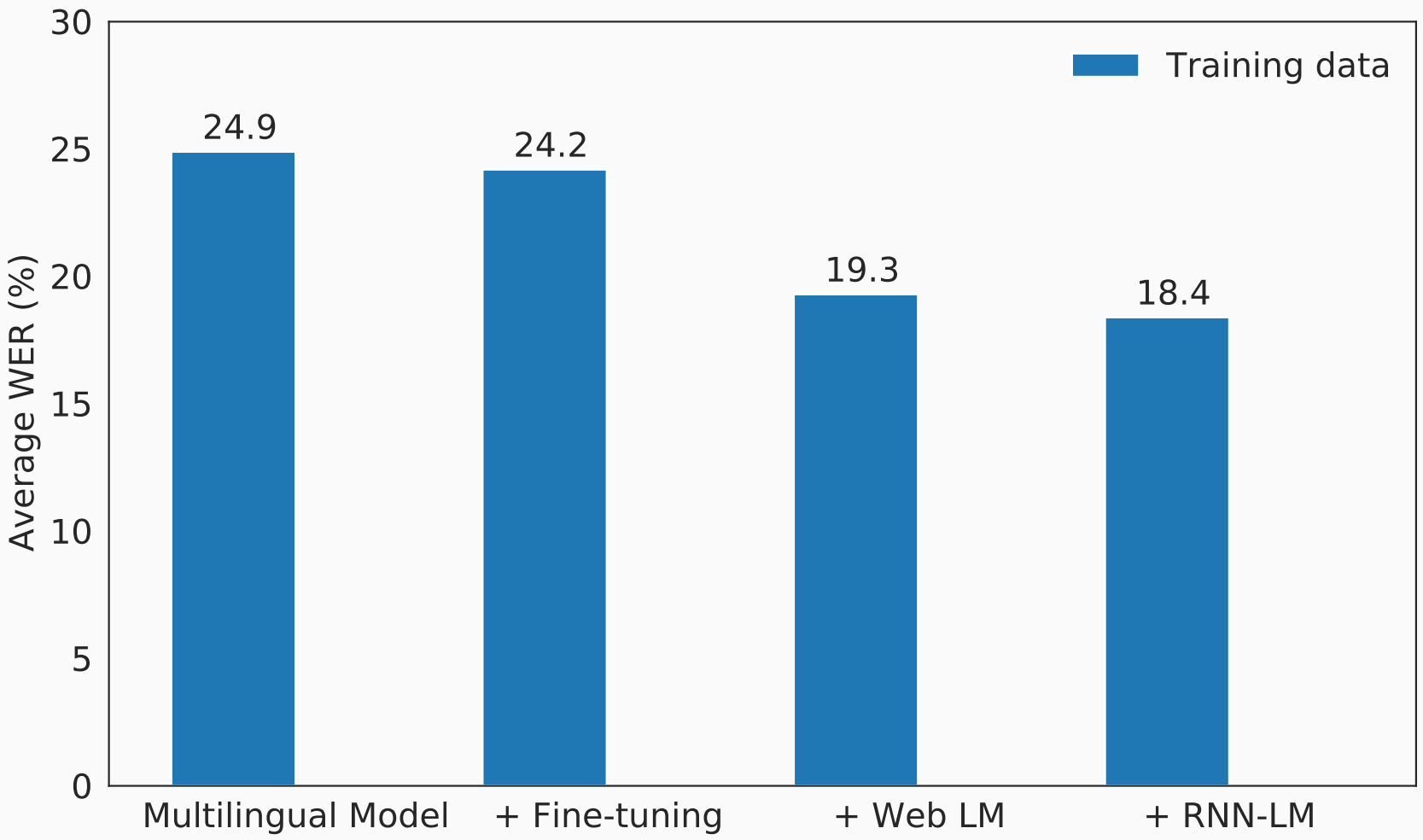
Results


30
25
20
15
10
5
0
Average WER (%)
Training data
24.9
24.2
Multilingual Model
+ Fine-tuning

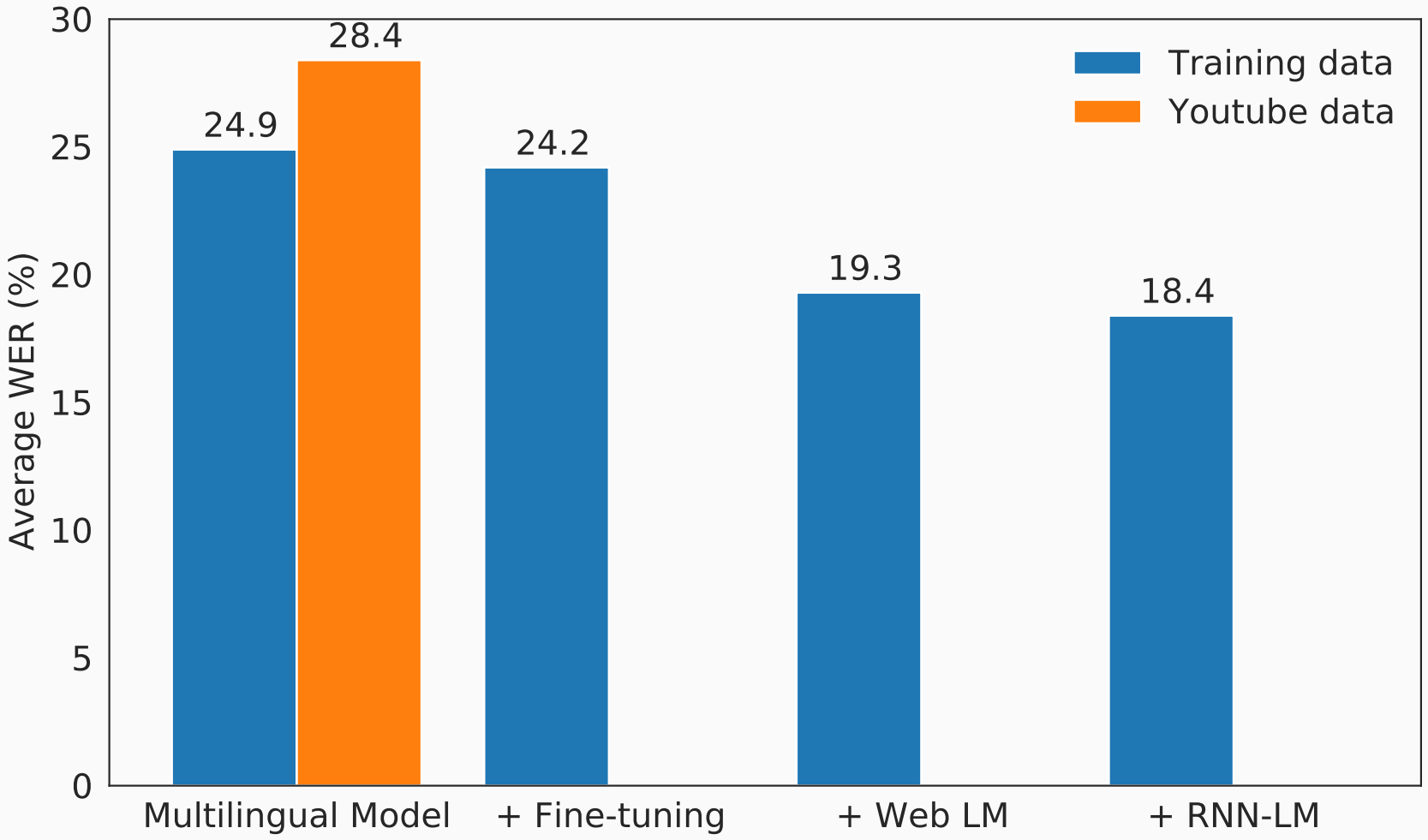
Results


30
25
20
15
10
5
0
Average WER (%)
Training data
24.9
24.2
19.3
Multilingual Model
+ Fine-tuning
+ Web LM

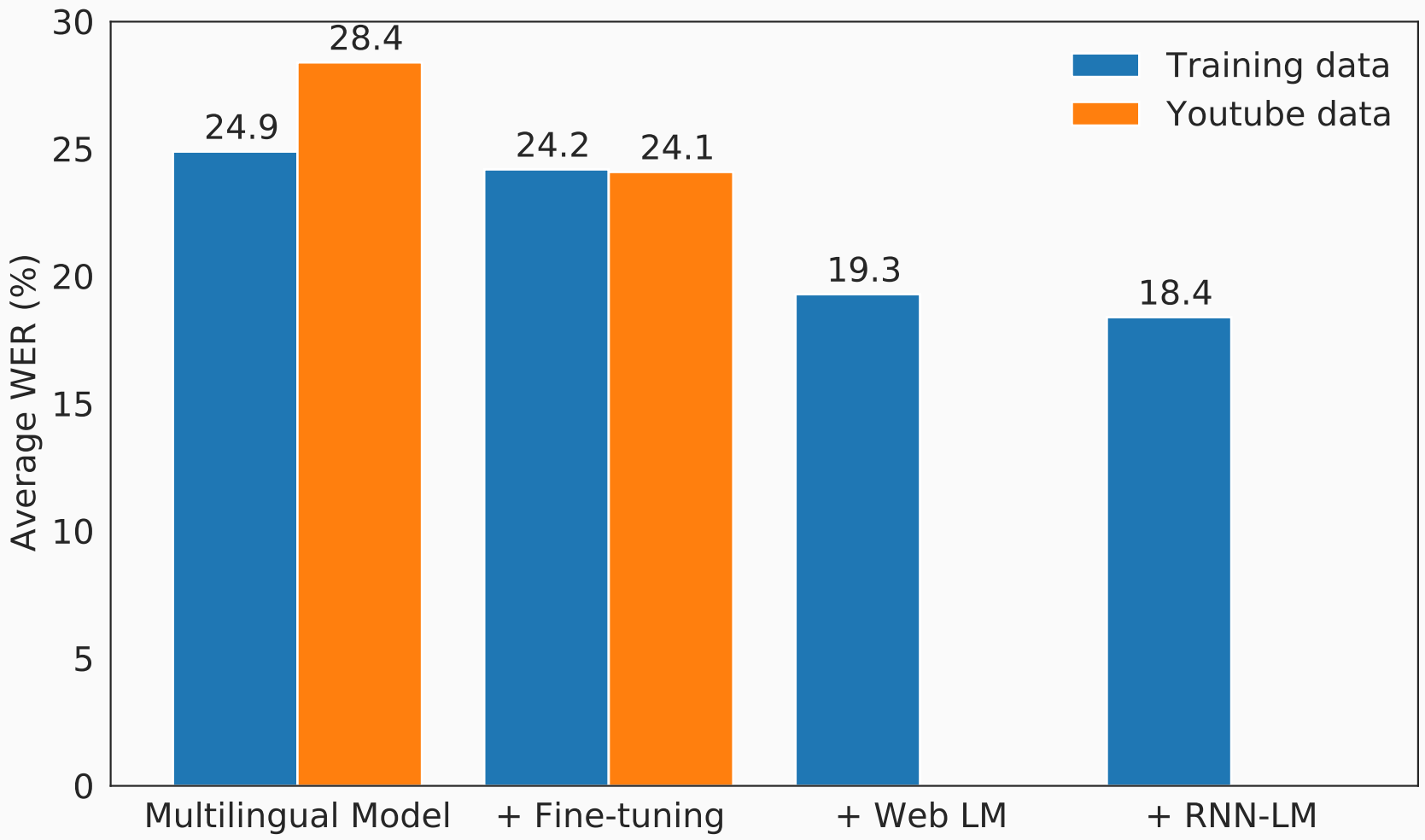
Results


Training data
Average WER (%)
30
25
20
15
10
5
0
24.9
24.2
19.3
18.4
Multilingual Model
+ Fine-tuning
+ Web LM
+ RNN-LM

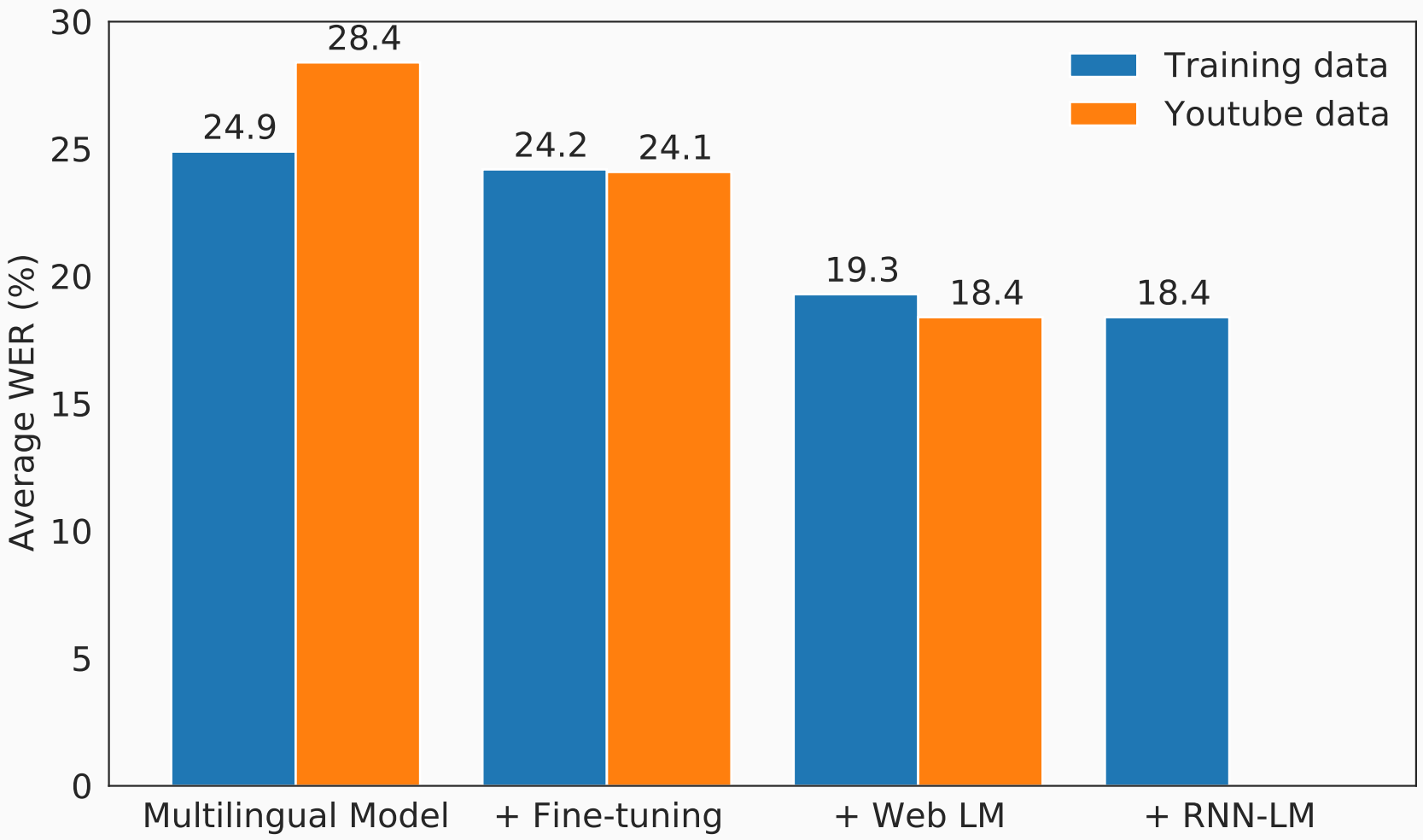
Results


30
25
20
15
10
5
0
Average WER (%)
24.9
28.4
24.2
19.3
18.4
Training data
Youtube data
Multilingual Model
+ Fine-tuning
+ Web LM
+ RNN-LM

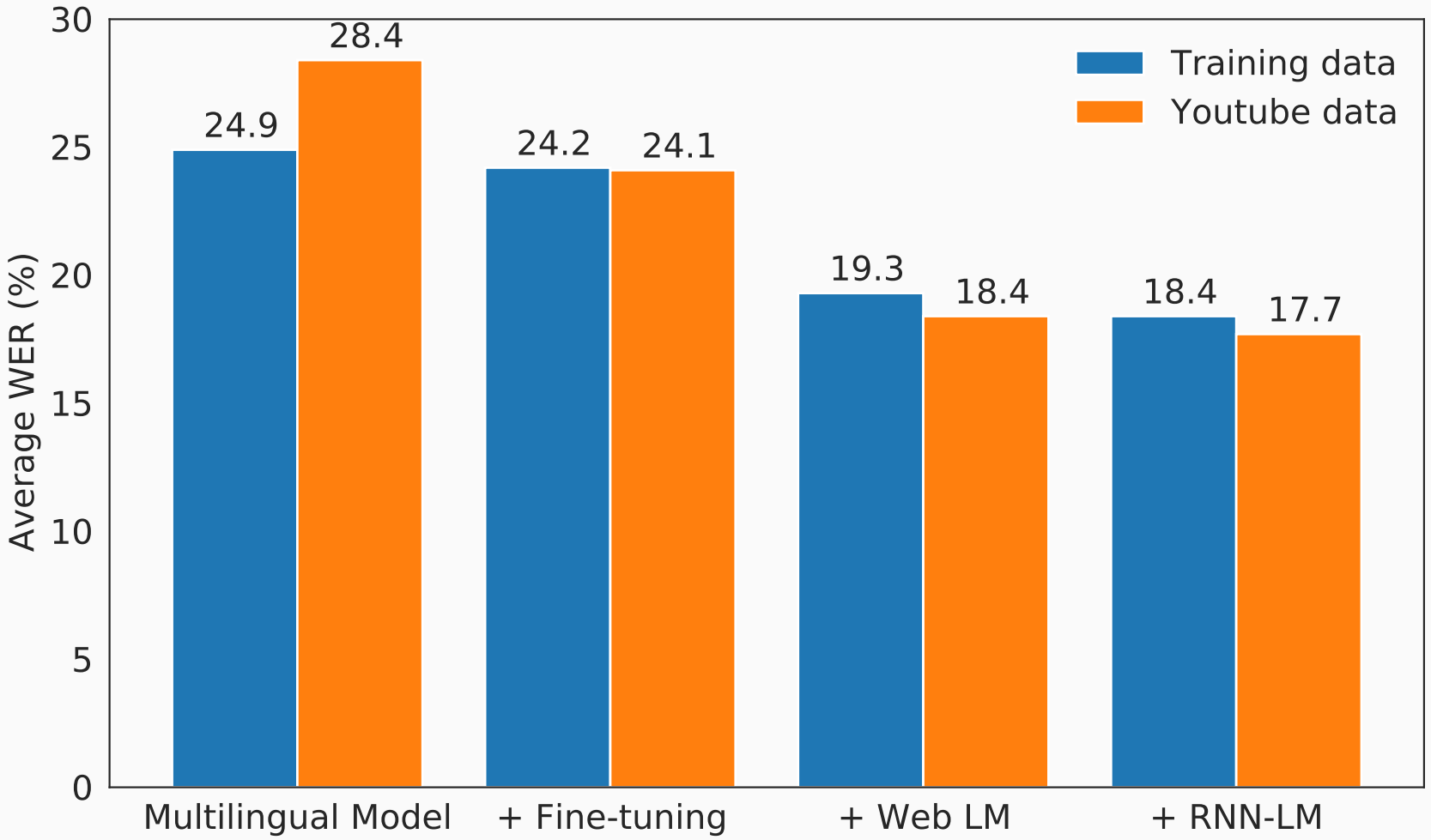
Results


Average WER (%)
30
25
20
15
10
5
0
24.9
28.4
24.2
24.1
19.3
18.4
Training data
Youtube data
Multilingual Model
+ Fine-tuning
+ Web LM
+ RNN-LM

Results


Average WER (%)
30
25
20
15
10
5
0
24.9
28.4
24.2
24.1
19.3
18.4
18.4
Training data
Youtube data
Multilingual Model
+ Fine-tuning
+ Web LM
+ RNN-LM

Results

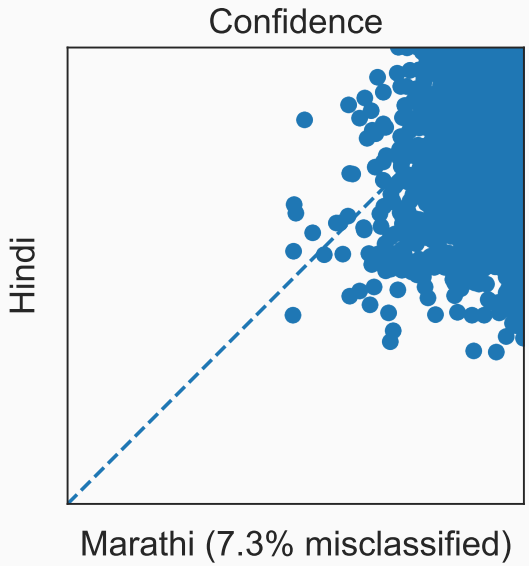

Average WER (%)
30
25
20
15
10
5
0
Training data
Youtube data
24.9
28.4
24.2
24.1
19.3
18.4
18.4
17.7
Multilingual Model
+ Fine-tuning
+ Web LM
+ RNN-LM

Language Identification

Blind test set does not contain language id, therefore we used confidence based language identification.

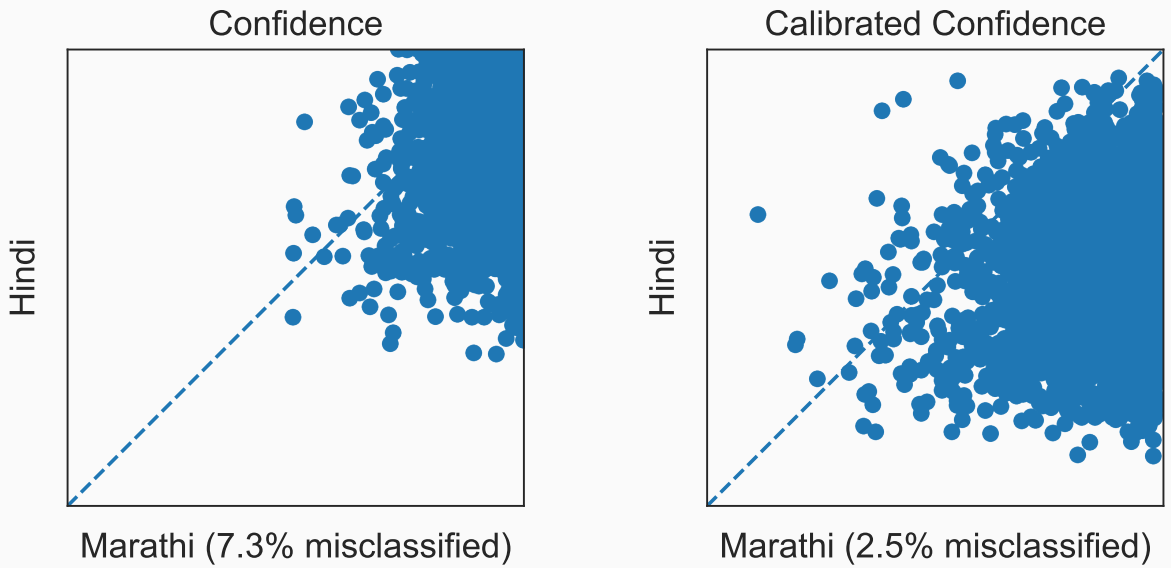
Language Identification

Blind test set does not contain language id, therefore we used confidence based language identification.



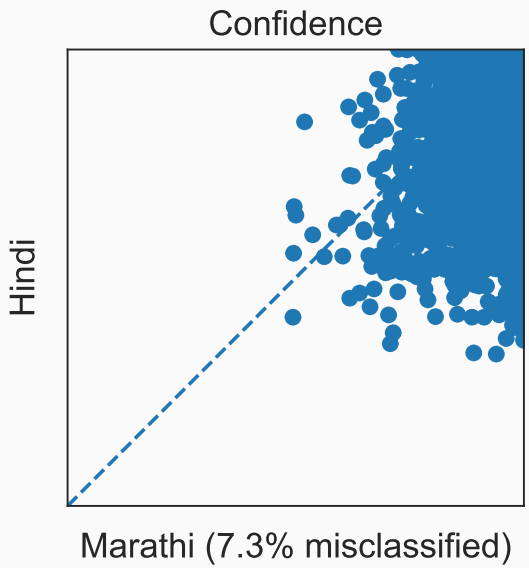
Language Identification

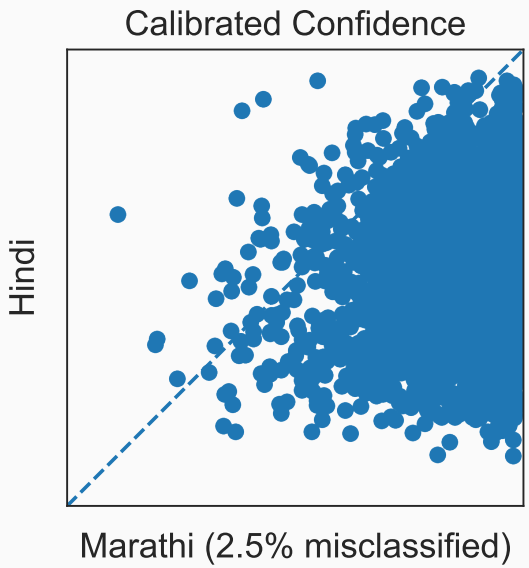
Blind test set does not contain language id, therefore we used confidence based language identification.

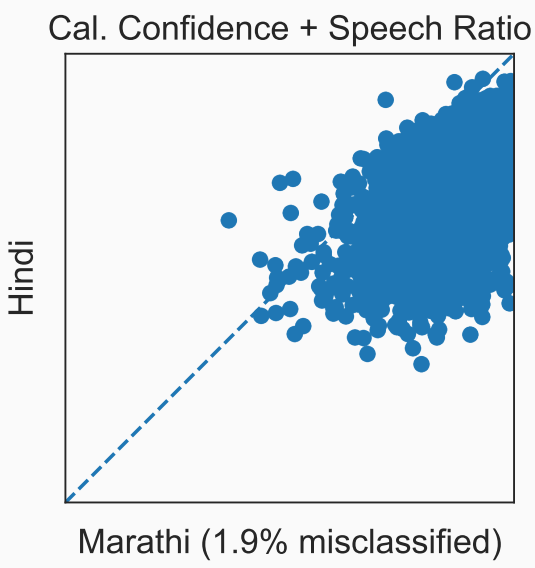


Language Identification

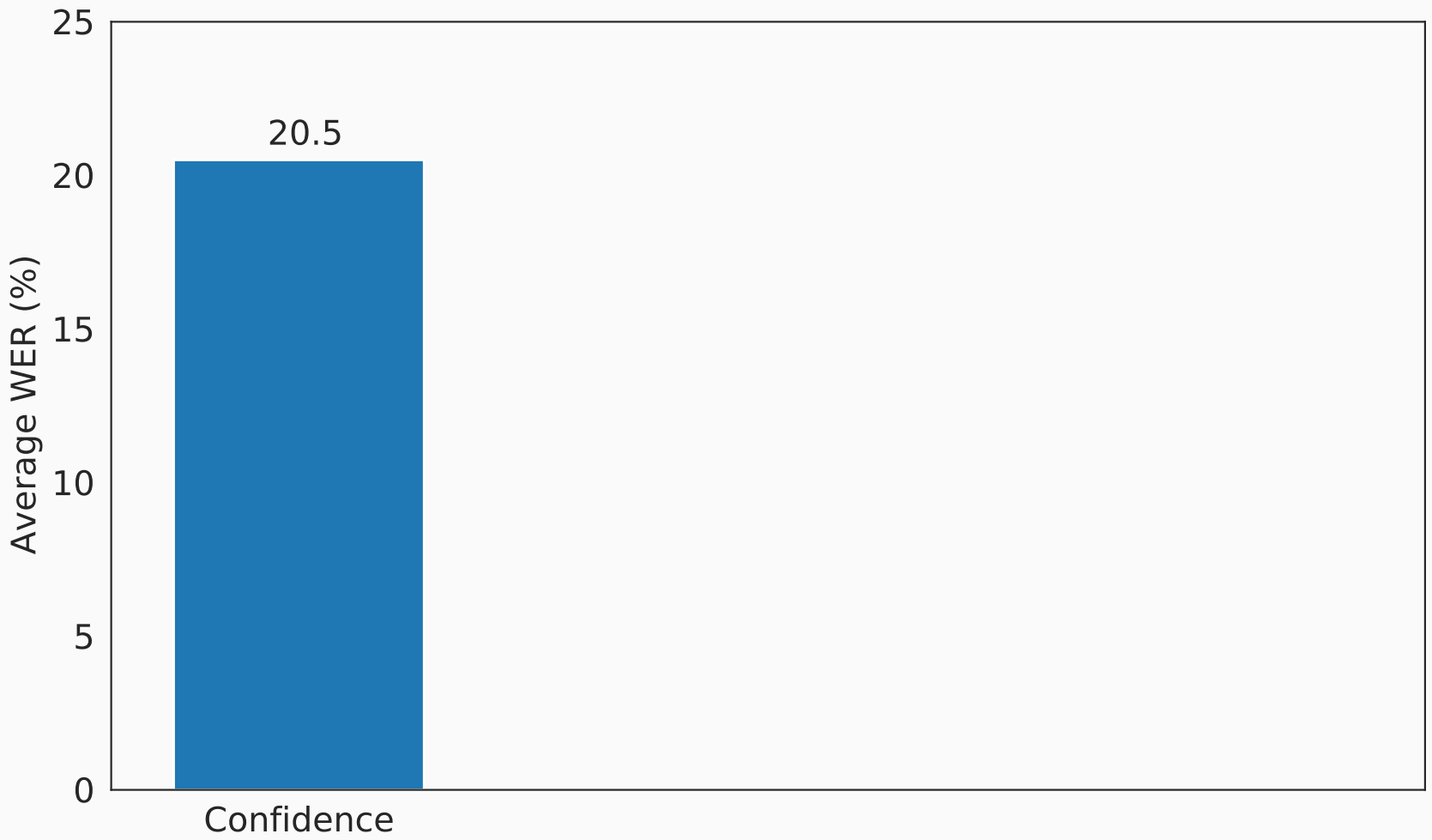
Blind test set does not contain language id, therefore we used confidence based language identification.



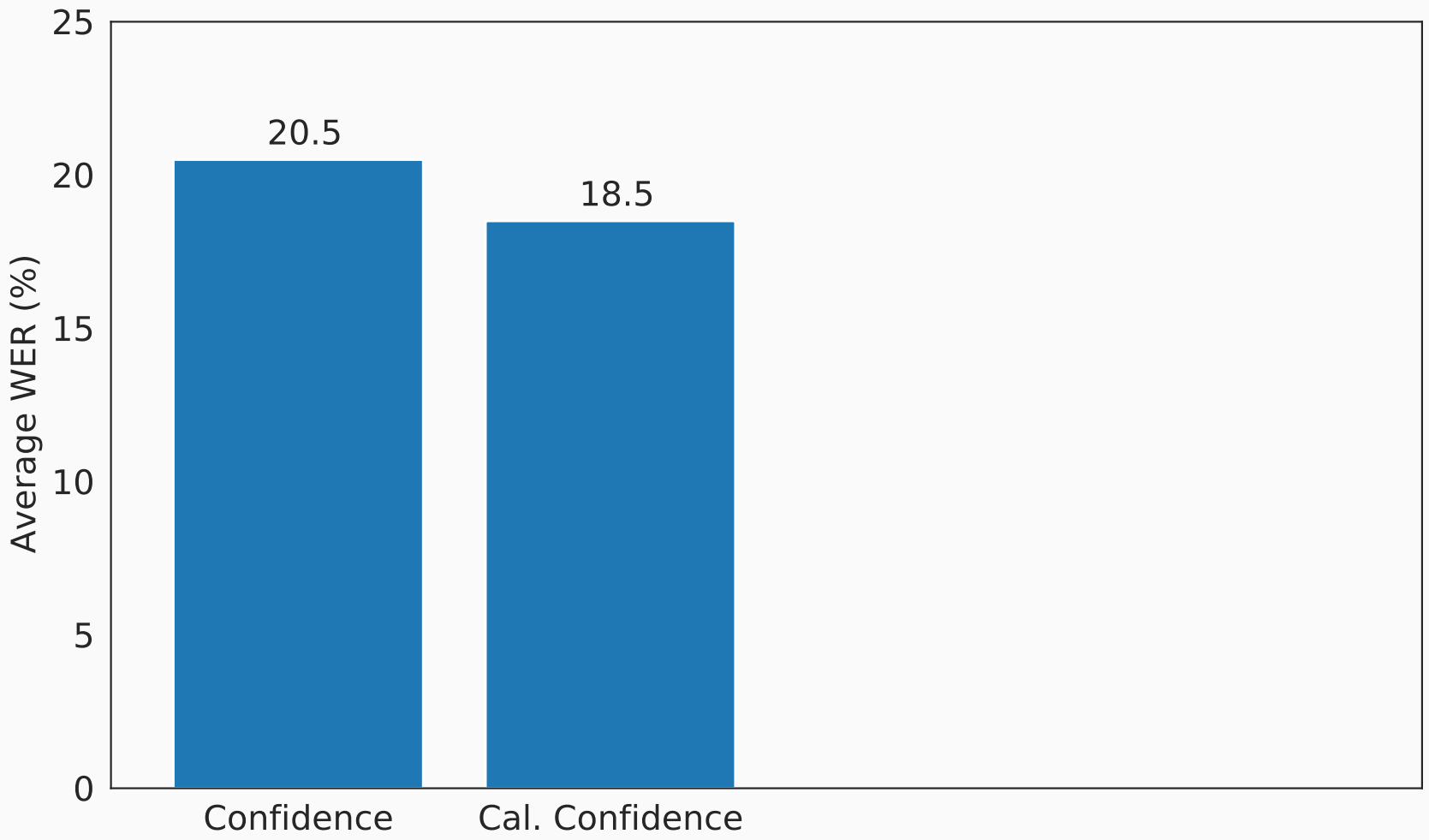




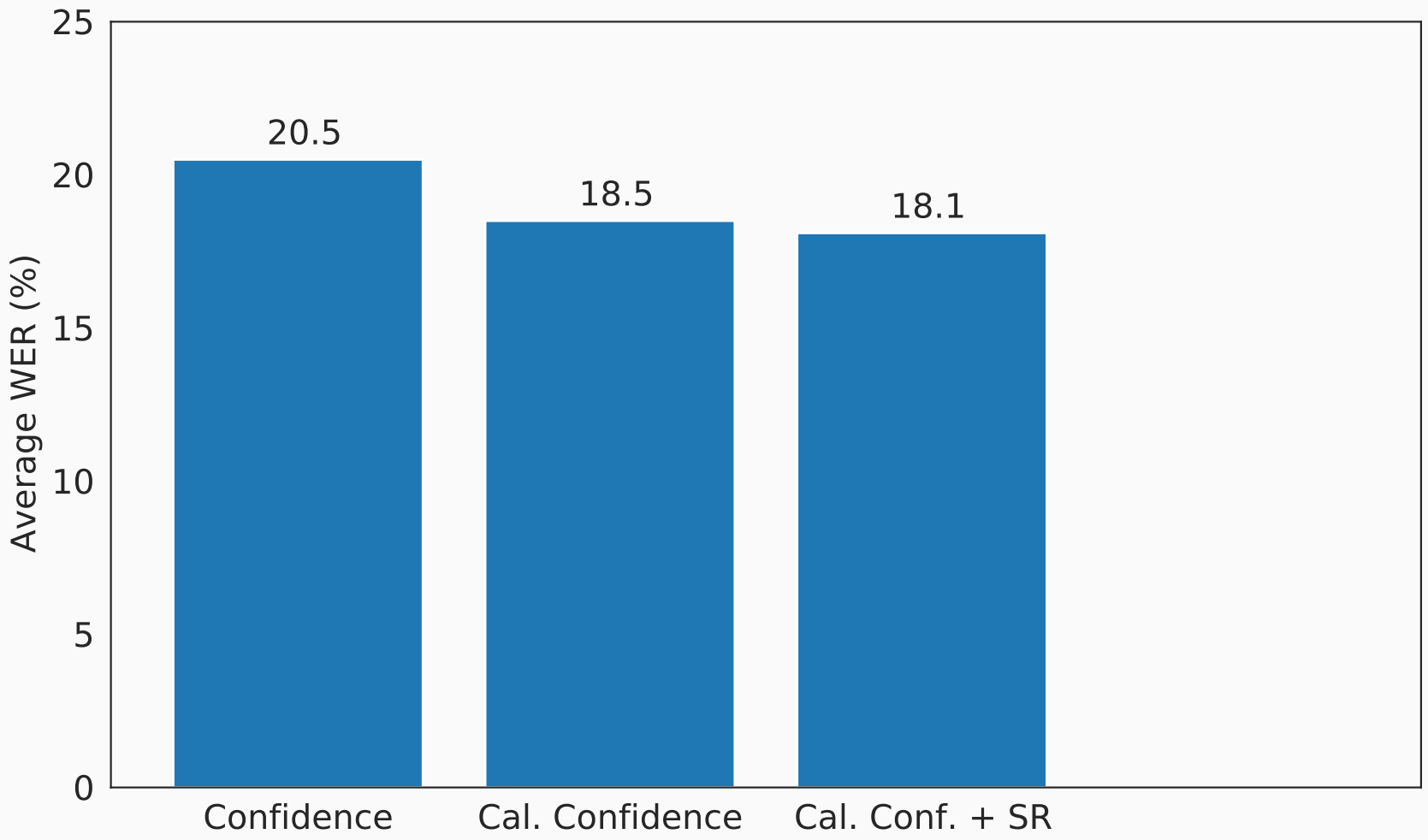
Language Identification Results



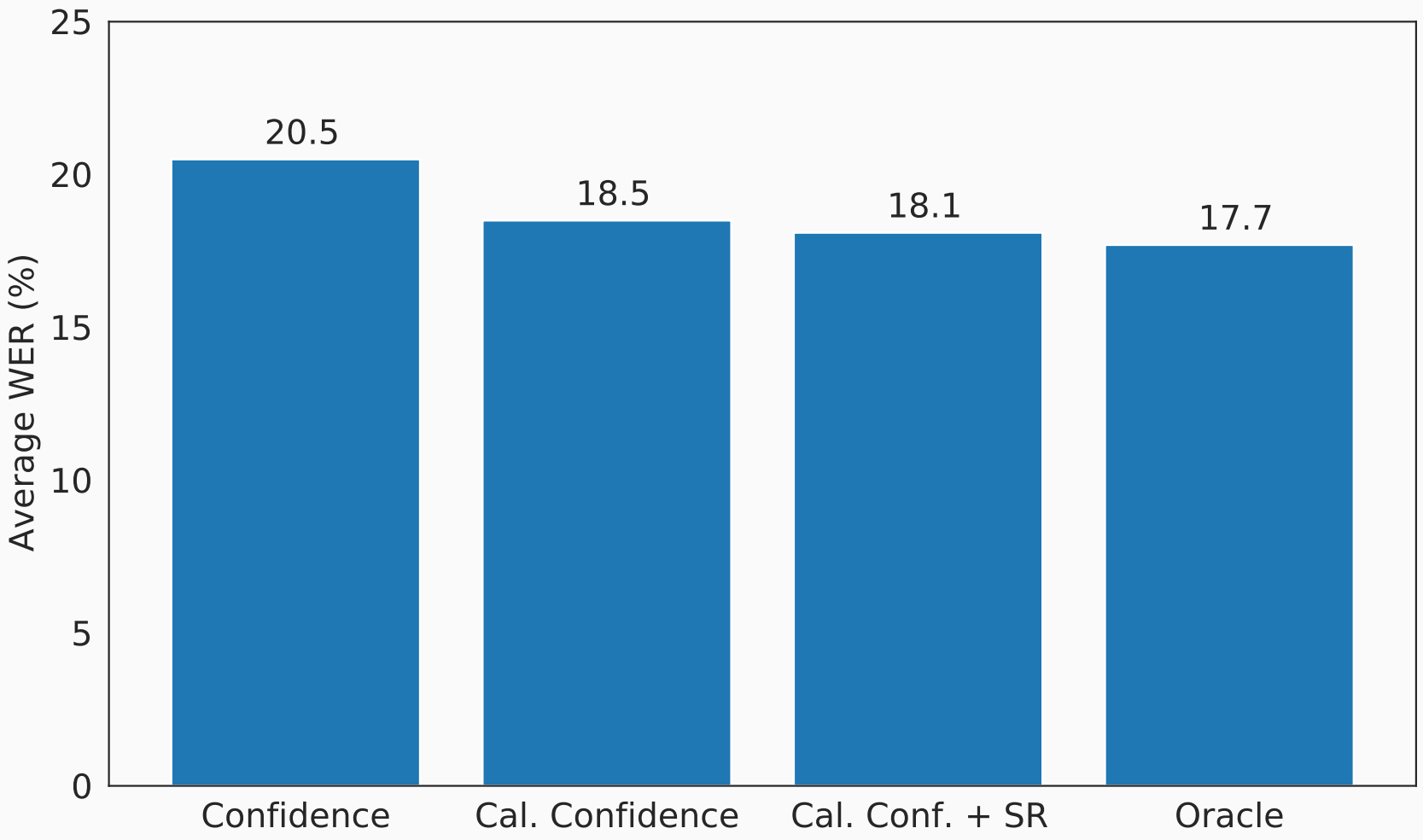
Language Identification Results



Language Identification Results



Language Identification Results



ASR models for low-resource languages can be trained with standard Kaldi recipes and crawled text/audio data.

ASR models for low-resource languages can be trained with standard Kaldi recipes and crawled text/audio data.

Confidence-based language identification works well, but is very expensive for deployment.

Subtask 2 System Overview

Acoustic Model

- CNN-TDNN with language specific outputs
- Multilingual training with LF-MMI

Language Model

- 3-gram, RNN-LM
- Data:
  - Training data
  - Hindi/Bengali CommonCrawl data
  - English SpokenTutorial.org subtitles

Lexicon

- Provided lexicon used language-specific units.

Lexicon

- Provided lexicon used language-specific units.

hello	HH AH L OW
world	W ER L D
नमस्ते	न म स ् त े
दुनिया	द ु न ि य ा

Lexicon

- Provided lexicon used language-specific units.

hello	HH AH L OW
world	W ER L D
नमस्ते	न म स ् त े
दुनिया	द ु न ि य ा

- To use the same units we trained a phone matcher using automatically crawled Wikipedia data.

Training the Phone Matcher

Training the Phone Matcher

David

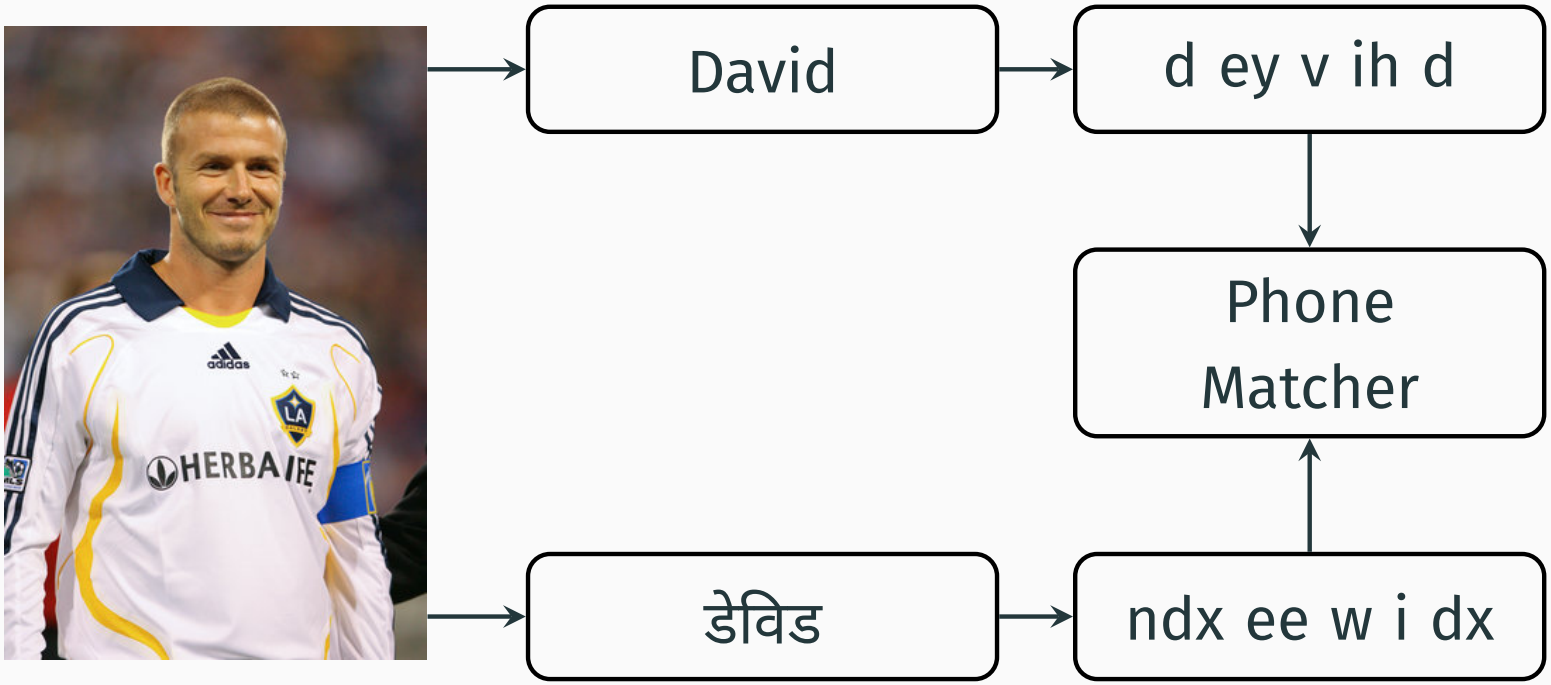
डेविड

Training the Phone Matcher

David → d ey v ih d

डेविड → ndx ee w i dx

Training the Phone Matcher



Language Modeling

The data contains a lot of English technical terms which might be rare in Hindi/Bengali CommonCrawl.

Language Modeling

The data contains a lot of English technical terms which might be rare in Hindi/Bengali CommonCrawl.

We downloaded subtitles of English SpokenTutorial.org videos, because SpokenTutorial.org was a common term in the training data.

Language Modeling

The data contains a lot of English technical terms which might be rare in Hindi/Bengali CommonCrawl.

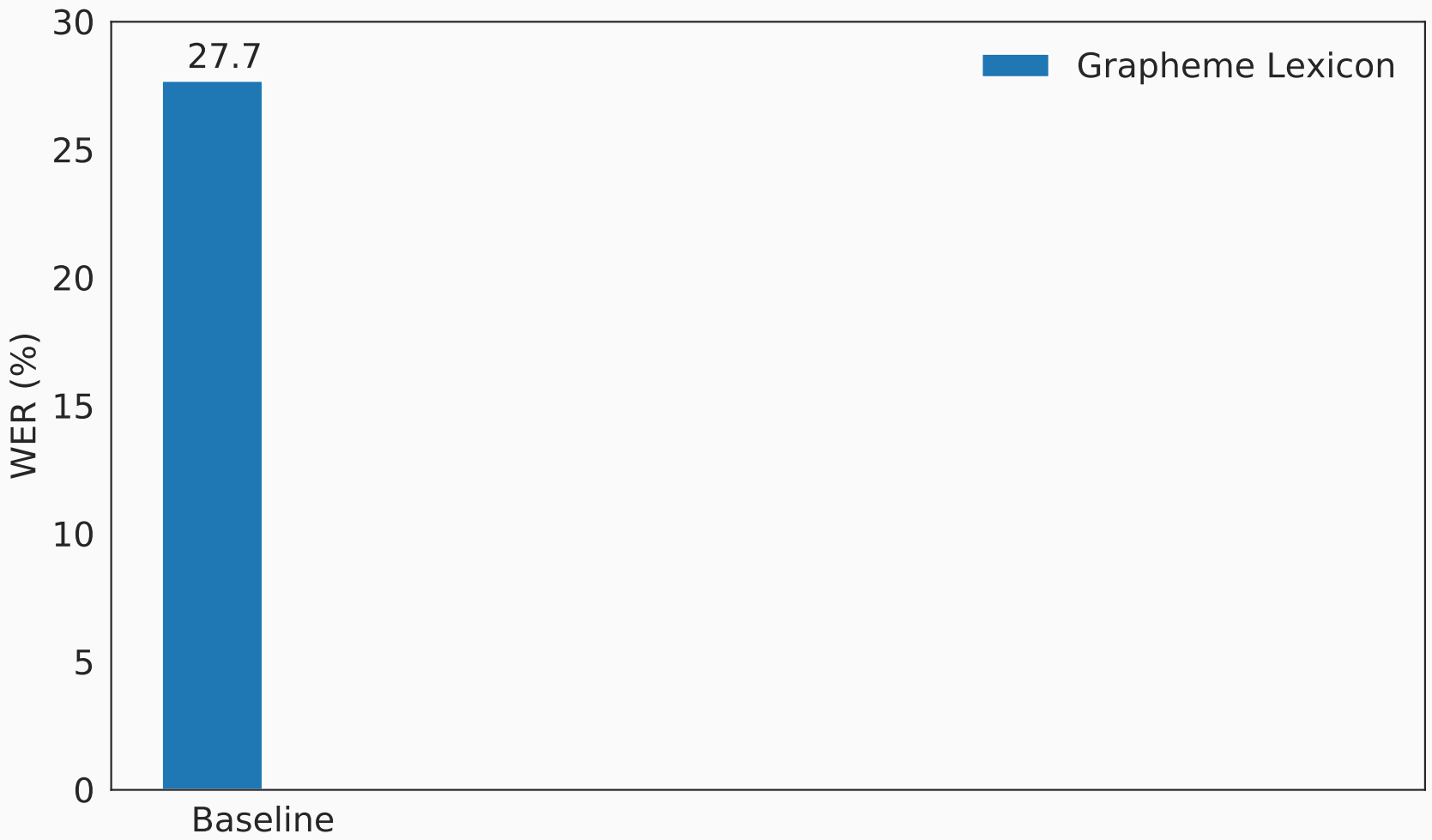
We downloaded subtitles of English SpokenTutorial.org videos, because SpokenTutorial.org was a common term in the training data.

The final language model was a mixture of language models trained on:

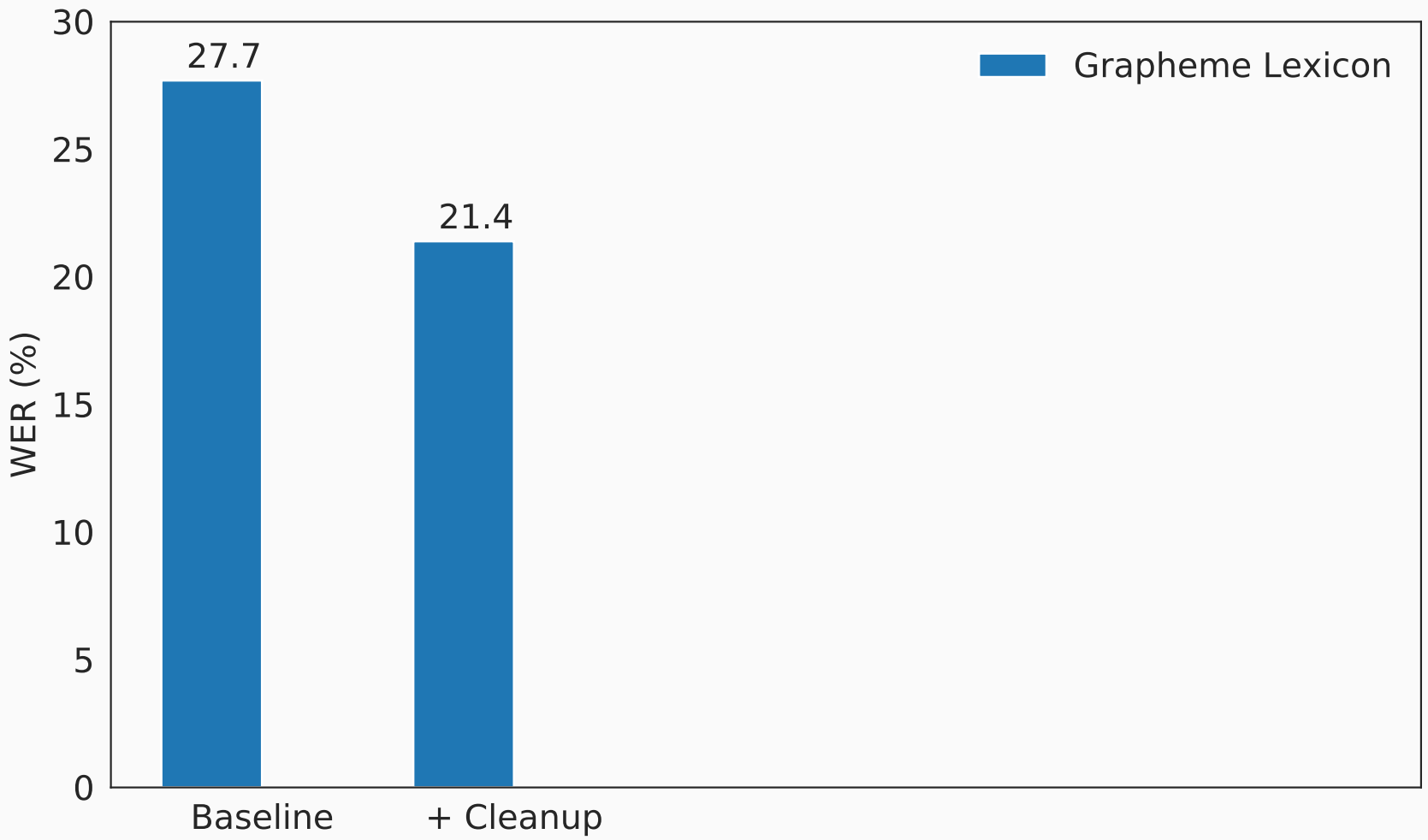
- Training data
- CommonCrawl data
- English SpokenTutorial.org subtitles

Interpolation weights were estimated on the dev data.

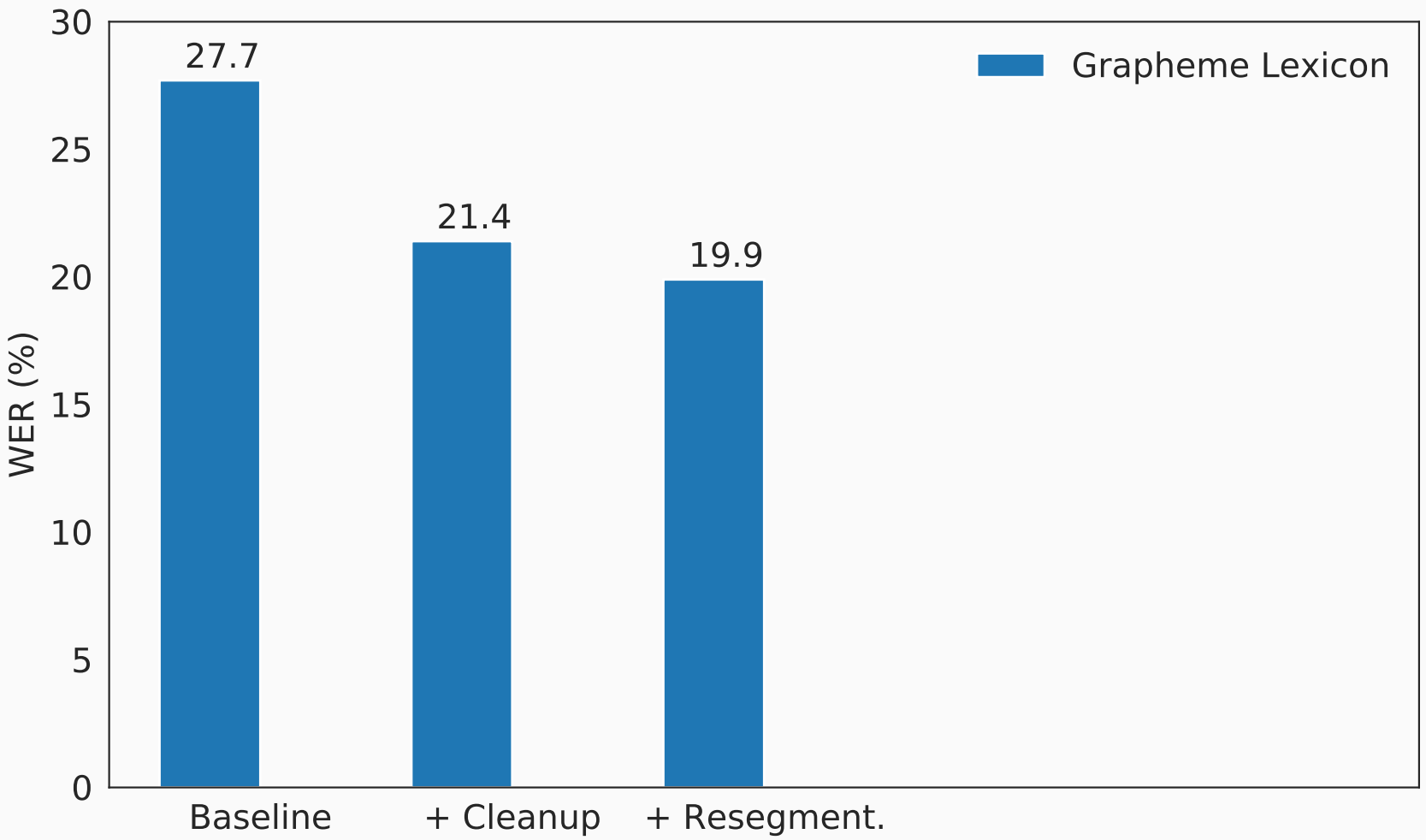
Hindi-English Results



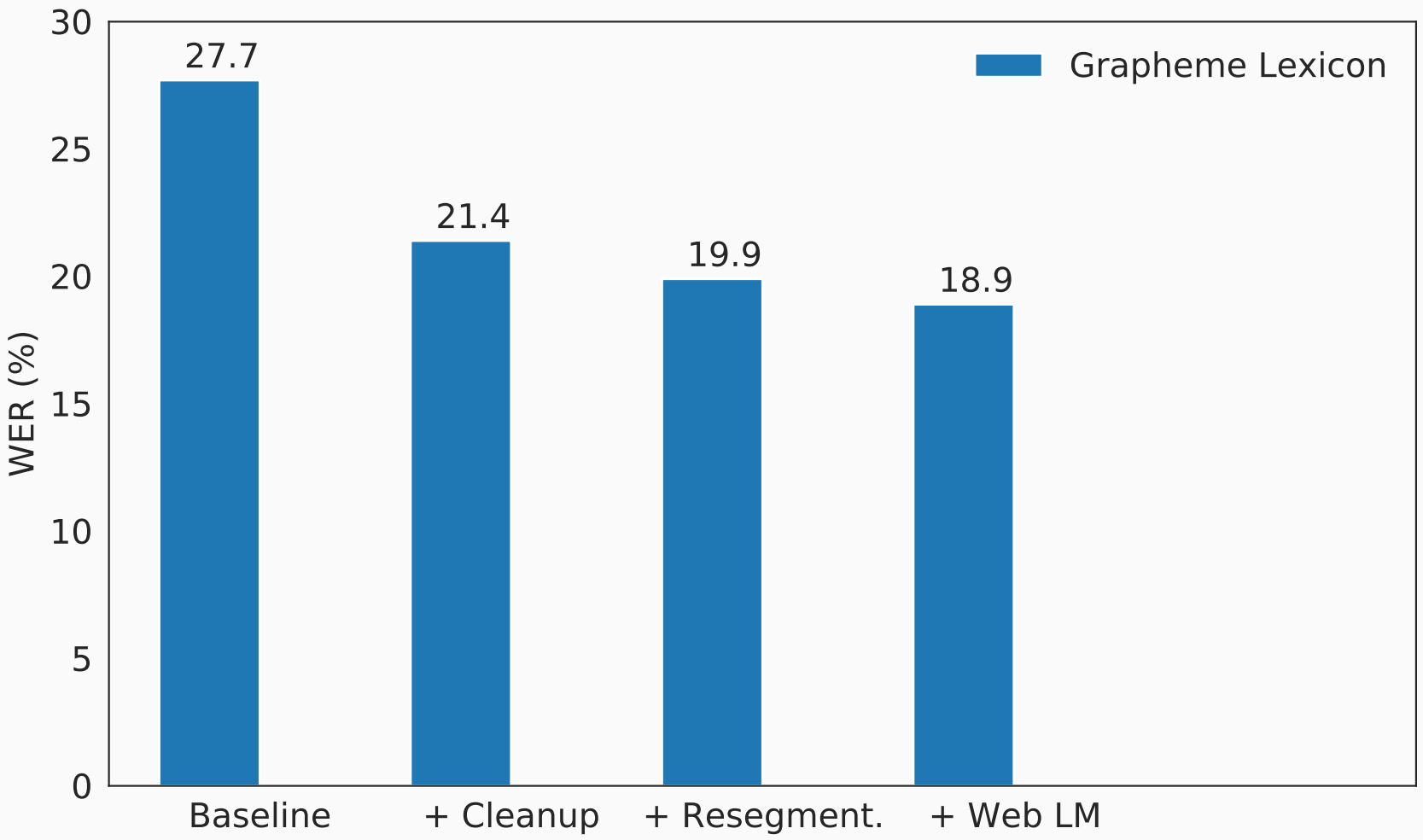
Hindi-English Results



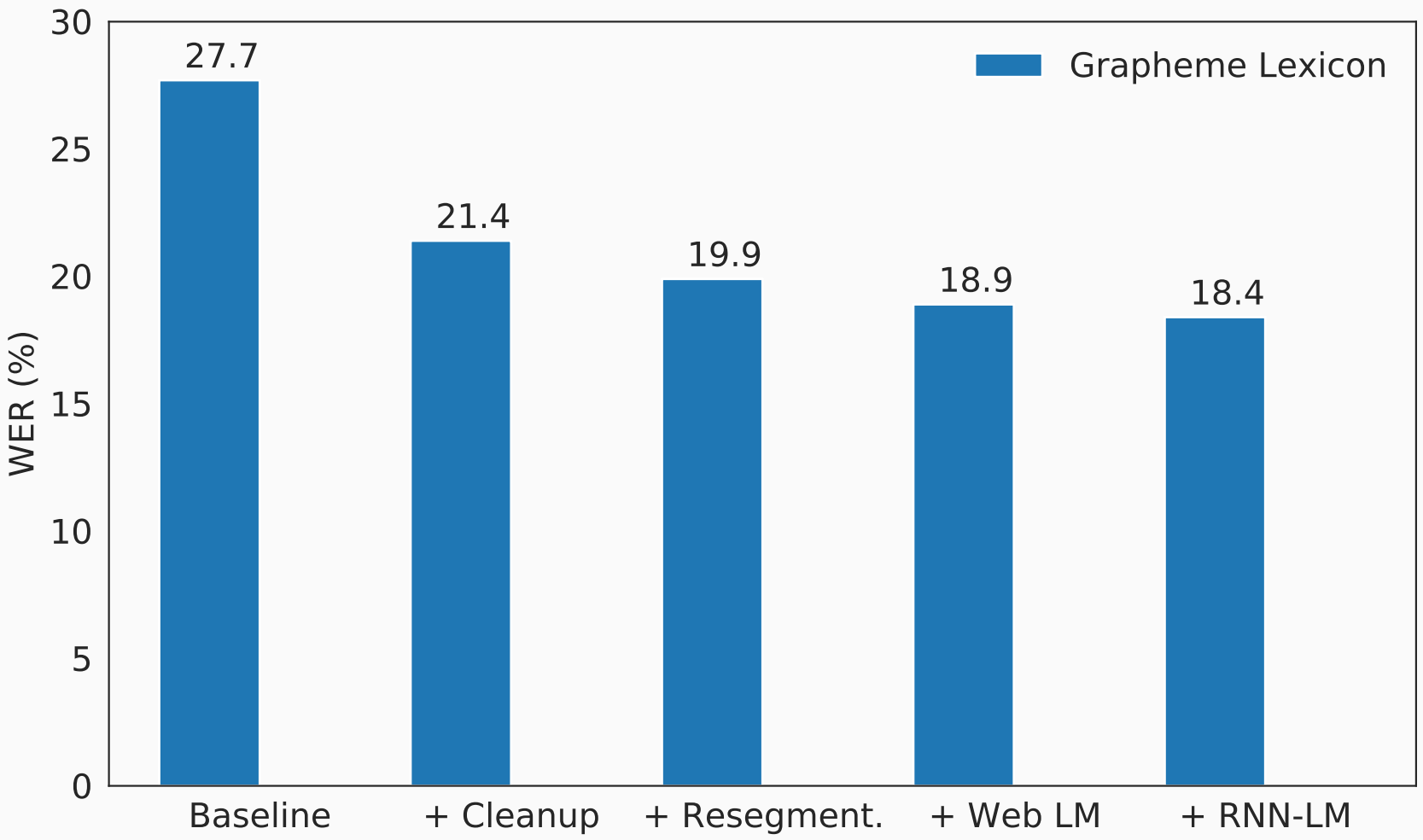
Hindi-English Results



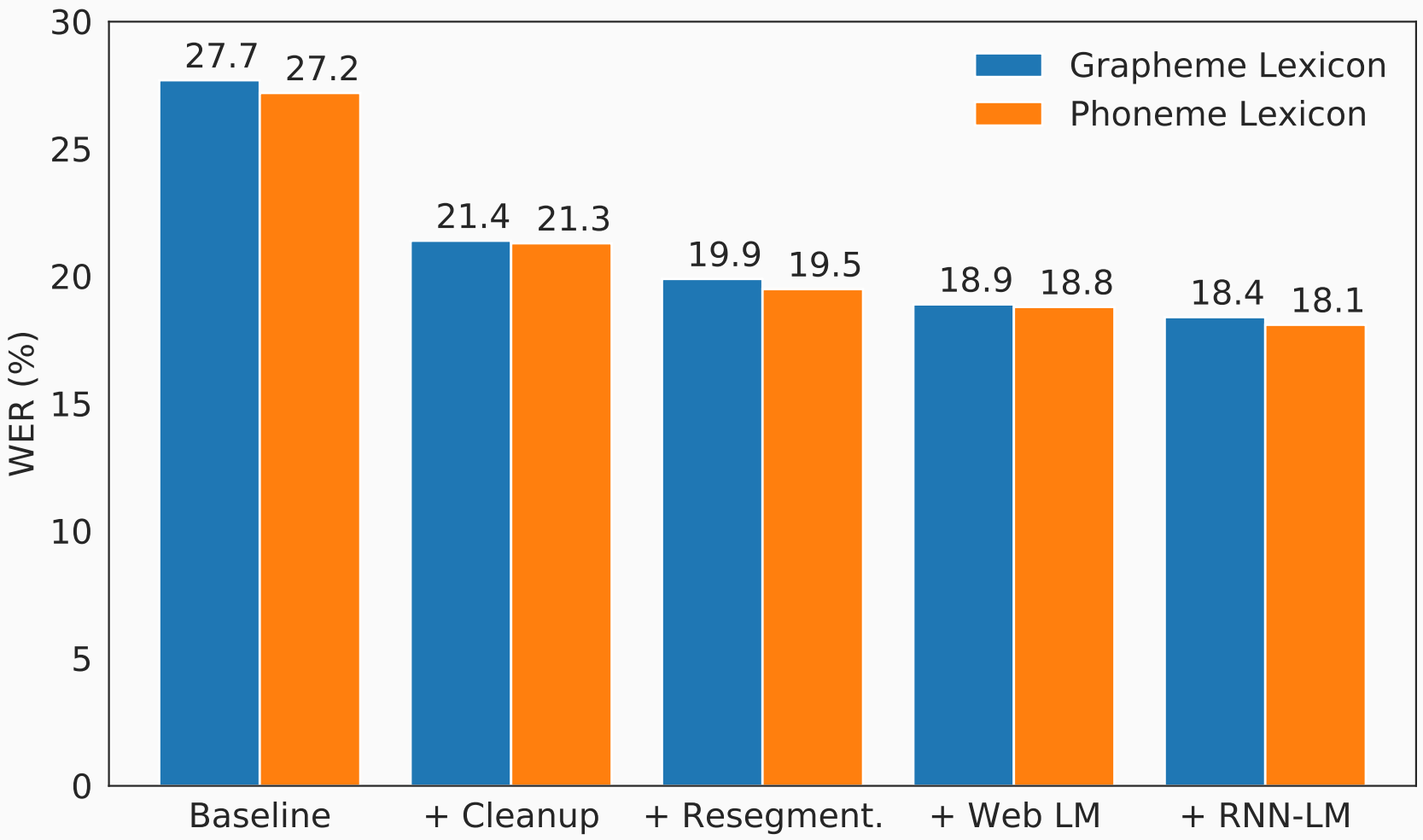
Hindi-English Results



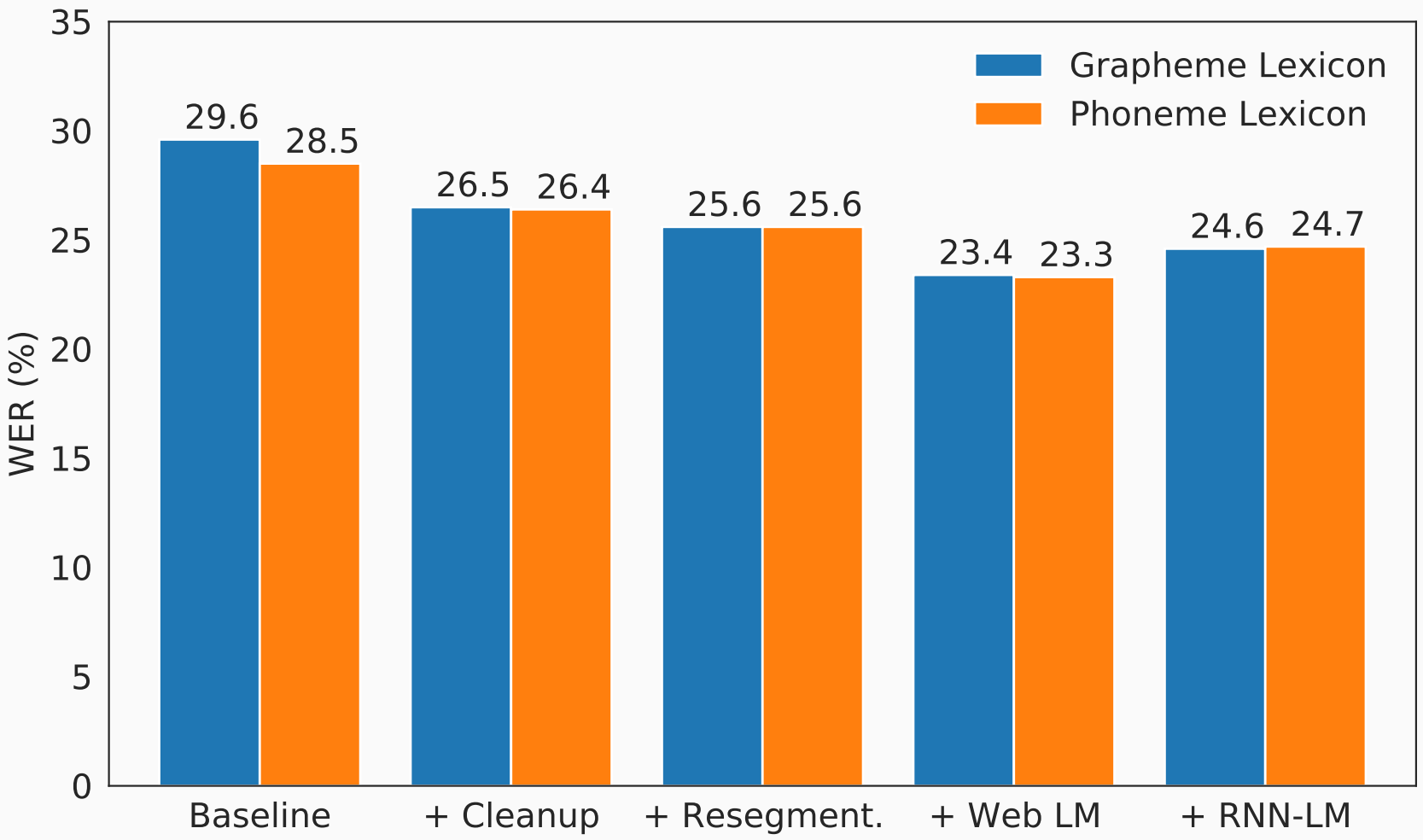
Hindi-English Results



Hindi-English Results



Bengali-English Results



Training data cleanup and appropriate LM data were very important for this challenge.

Training data cleanup and appropriate LM data were very important for this challenge.

Phone matching can be learned automatically using Wikipedia data for languages using different scripts.

Training data cleanup and appropriate LM data were very important for this challenge.

Phone matching can be learned automatically using Wikipedia data for languages using different scripts.

It is not clear that naively mixing language models would work in more challenging code-switching conditions.