This work has been supported by the DFG under contract no. Ha3455/12-1.

Interspeech 2017, Stockholm

Necessary to find words not in initialization vocabulary

- Measures: Segmentation F-score, Word Error Rate (WER)

Given fraction of text corpus randomly chosen to initialize word segmentation

Language Model

Leveraging Text Data for Word Segmentation for Underresourced Languages

Word Segmentation Extension

Experimental Setup

Different experiments:

- AUD: System from [1] used
- AUD on test set, AUD is trained semi-supervisedly on AUD on Xitsonga Database, NCHLT Xitsonga text corpora (40190 sentences) [5]

Conclusion

Results

References