Leveraging Text Data for Word Segmentation for Underresourced Languages

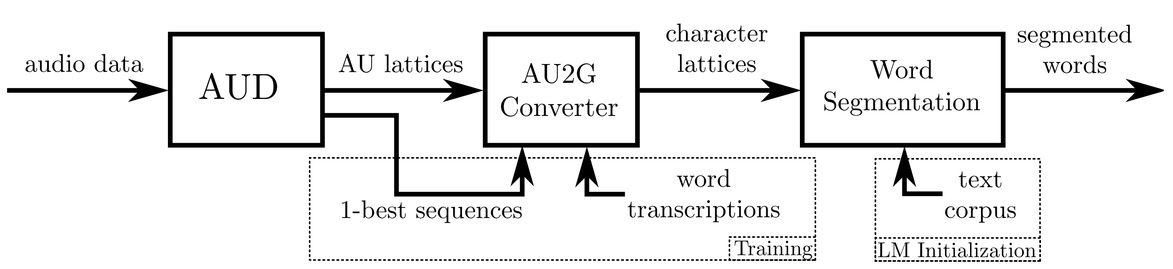
Thomas Glarner[1], Benedikt Boenninghoff[2], Oliver Walter[1], Reinhold Haeb-Umbach[1],
[1]Paderborn University, Germany [2]Ruhr-Universität Bochum, Germany

{glarner, walter, haeb}@nt.uni-paderborn.de, http://nt.uni-paderborn.de
benedikt.boenninghoff@rub.de, https://www.ruhr-uni-bochum.de/ika/

Introduction

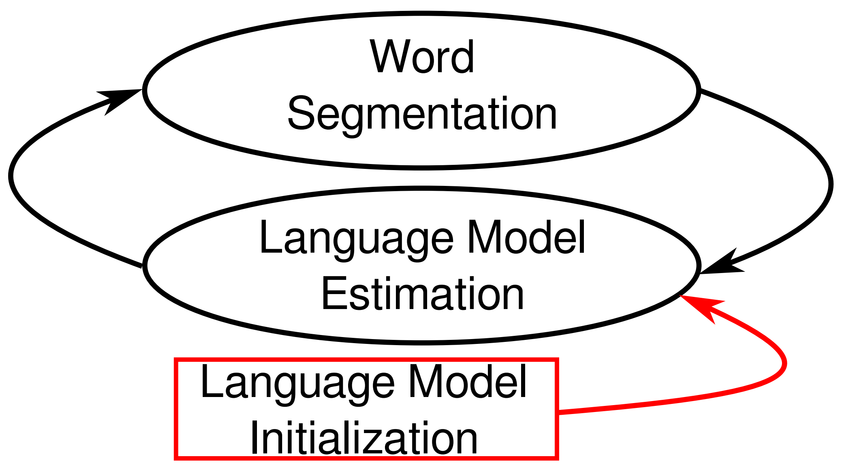
- Common situation in speech recognition for underresourced languages:
  - No pronunciation dictionary or detailed annotations
  - unknown phonemic inventory
  - Word-level transcript available for some speech recordings
  - Some unrelated textual data available
- How to connect automatically learned phonemic transcripts and text?
  → Train semisupervised Acoustic Unit-to-Grapheme converter
- How to take advantage of disjoint text data?
  → Language model initialization for word segmentation

Proposed System



- AUD: System from [1] used

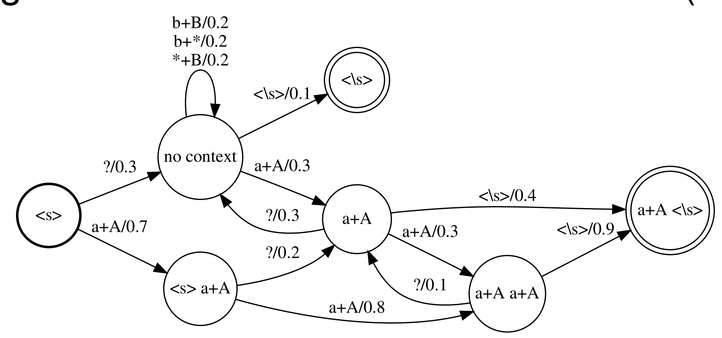
Word Segmentation Extension



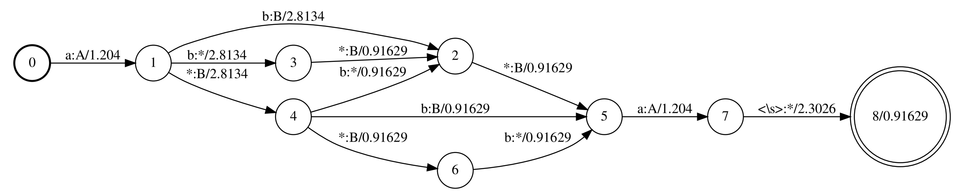
- System from [2]
- Added ability for Language Model initialization
- Necessary to find words not in initalization vocabulary

AU-to-Grapheme Conversion

- Use of Sequitur G2P [3]
- supervised training on subset of utterances
- Problem to solve: Word boundaries unknown
- Usage Principle similar to [4]:
  - Trained **per-utterance on unaligned word-level transcription** and 1-best label sequence from AU lattice
  - Extract highest-order LM and convert to WFST (GLM)



  - Build AU-to-Graphone and Graphone-to-Letter WFSTs.
  - Compose $l = \text{Prune}(a \circ A2G \circ_\phi GLM \circ G2L)$.



Experimental Setup

- Different experiments:
  - Kaldi Phoneme Recognizer on WSJ Database, WSJ LM training corpus (1631456 sentences)
  - AUD on WSJ Database, text corpus as above
  - AUD on Xitsonga Database, NCHLT Xitsonga text corpora (40190 sentences)[5]
- Split in training and test (WSJ: 5392/5391 utt., Xitsonga: 2029 utt. each)
- AUD on test set, AU2G is trained semi-supervisedly on training set, segmentation done on test set
- Given fraction of text corpus randomly chosen to initialize word segmentation LM
- Measures: Segmentation F-score, Word Error Rate (WER)

Results

Size of LM init corpus		10%	1%	0.1%	0.01%	0%
WSJ Phn Recog.	F-score	79.8	78.8	75.0	63.1	51.8
	WER	24.9	25.9	30.6	46.5	61.3
WSJ AUD	F-score	28.3	28.2	26.8	22.7	13.6
	WER	77.5	77.5	78.7	83.1	92.5
		100%	10%	1%	0.1%	0%
Tso AUD	F-score	42.9	41.7	39.5	31.3	17.4
	WER	71.5	76.4	80.8	94.2	140.2

Conclusion

- Successful AU2G training possible without word boundaries
- word segmentation necessary to find unknown words
- Text data greatly improves performance of speech recognition
- Exemplary: Usage of 400 **Xitsonga** sentences (1%) leads to a relative WER improvement of 42.4%.

References

L. Ondel, L. Burget, and J. Cernocky, "Variational inference for acoustic unit discovery," in *Proceedings of the 5th Workshop on Spoken Language Technologies for Under-resourced languages*, vol. 81, 2016, pp. 80 – 86.

J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices," in *39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, may 2014.

M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Commun.*, vol. 50, no. 5, pp. 434–451, 2008.

M. Hannemann, Y. Trmal, L. Ondel, S. Kesiraju, and L. Burget, "Bayesian joint-sequence models for grapheme-to-phoneme conversion," in *42th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, 2017.

E. Barnard, M. H. Davel, C. J. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the south african languages," in *4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014.* ISCA, 2014, pp. 194–200.

Interspeech 2017, Stockholm

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