Unsupervised Word Discovery from Phonetic Input Using Nested Pitman-Yor Language Modeling

Oliver Walter*, Reinhold Haeb-Umbach*, Sourish Chaudhuri** and Bhiksha Raj**
*Department of Communications Engineering - University of Paderborn
**Language Technologies Institute - Carnegie Mellon University

Unsupervised language acquisition

- "Learn a language like a child"
- Two-level hierarchical process:
  - 1. acoustic unit (phone) discovery
  - 2. lexical unit (word) discovery

Here: Word discovery

- Input: character or phone sequence
  theabilitytolearnwithoutbeingprogrammed
- Output: segmented sequence (Words)
  the ability to learn without being programmed
- Unsupervised word discovery
- no initial lexicon and language model

Segmentation Algorithm

- Unsupervised segmentation algorithm
- Lexicon and language model learned
- Pitman-Yor language model
- Gibbs sampling based

Experiments

- Wall Street Journal data
- Error free phonemic transcriptions
- Performance on running text:
  - 57.1% recall of segmentations
  - 73.3% precision in segmentations