Effects of Adding Word Embeddings to Neural-Network-based Pitch Accent Detection

Abstract

Motivation
- PAD benefits from adding information from text: parts of speech, function vs. content words, word identity
- state-of-the-art deep learning methods use word embeddings to represent syntactic and semantic properties of words
- not previously used for PAD on transcribed speech

Findings
- word embeddings help most when word overlap is significant
- this tends to lead to overfitting → generalization challenging

Model

Required input data
- acoustic signal (.WAV) and transcriptions
- time-aligned at the word level

Convolutional Neural Network
- input matrix: frame-based acoustic features for each trigram
- position features indicate current word
- 2-layer convolutional neural network
  - 1st layer: 100 kernels, size 6 x 7
  - 2nd layer: 100 kernels, size 4 x 2
- dropout: p = 0.2, /2 regularization

Acoustic Features
6 low-level descriptors extracted using OpenSMILE [1]
- RMS energy*, loudness*, smoothed F0, voicing probability, harmonics-to-noise-ratio, zero-crossing rate

Feed-forward Network and Word Embeddings
- input: for each unigram or word in trigram
- 300-dimensional word embedding vector
- pre-trained word embeddings: word2vec [2], GloVe [3]
- used as non-trainable matrix weights in hidden layer
- dropout p = 0.8, /2 regularization
- bottleneck with variable size n

Data

- Boston University Radio News Corpus [4]
  - 27k words, 51.5% accented
- Boston Directions Corpus (read & spontaneous) [5]
  - 19k words, 55.5% accented
- LeaP corpus of non-native speech (read & retold stories) [6]
  - 15k words, 43.1% accented

Out-of-vocabulary words and performance on stopwords

- word2vec omits stopwords a, and, of, to
- OOVs represented as vector of ones

Results

- All results shown in accuracy (%) averaged using 10-fold crossvalidation and 5 repetitions.
- Left: within-corpus and cross-corpus experiments using GloVe unigram embeddings, n = 10
- Below: within-corpus experiments using embeddings with and without context and varying bottleneck sizes

Acknowledgements

This work was funded by the German Research Foundation DFG (SFB 732, A8).

References