Dialog Act Classification

using Word Embeddings & Acoustic Features

Jens Beck, Fabian Fey, Richard Kollotzek
Institute for Natural Language Processing, University of Stuttgart

Task Introduction

- Dialog Act Classification describes the process of automatically predicting the dialog act of the current speech and textual information
- We present an approach using a Convolutional Neural Network (CNN) which classifies utterances in four classes:
  - statement (I think I read about that in the paper)
  - question (Well where do you take those things)
  - opinion (It was really good)
  - backchannel (Uh-huh)
- We combine two different inputs:
  - Lexical features
  - Acoustic features

Data

Switchboard

- We use a subset of the Switchboard Telephone Speech Corpus which consists of lexical and acoustic data
- Our subset includes 40,556 sentences
- The lexical dataset is divided in training, development and test data
- The acoustic dataset includes a recording for every utterance

Dataset\Channel	opinion	question	backchannel	statement	Sum
training	4984	2150	6792	14459	28385
development	1068	460	1455	3098	6081
test	1070	463	1458	3099	6090

MFCC features

- With OpenSmile we extract the MFCC features for every sentence
- The MFCC features are extracted every 10ms with a frame size of 25ms
- This results in 13 features for each measurement point

word2vec

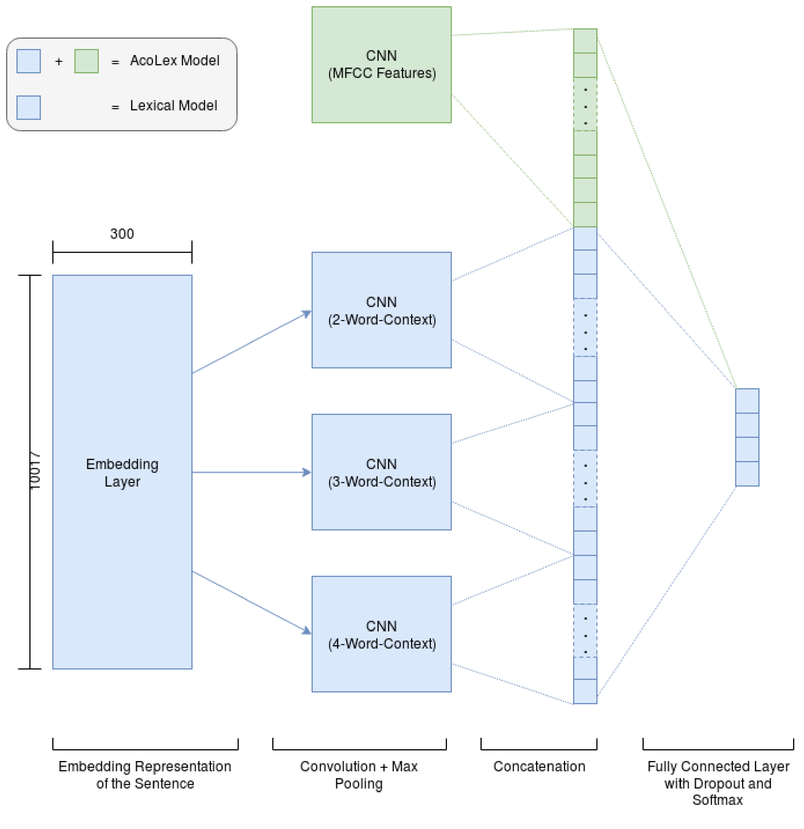
- For the word embedding layer we use the pre-trained Google word2vector model
- It contains 3 million words and phrases with a 300-dimensional vector each

Data Preprocessing

1. All words from the three datasets are indexed in a word list for use in the embedding matrix
2. Each word from the training set is given it's corresponding vector from the word2vec model
3. A random vector is assigned if the word is not in the word2vec model or from the test and development set
4. Each sentence is converted to a sequence with the corresponding indexes from the word list
5. The maximum sentence length is set to 100 words if the sentence is shorter the remaining space is zeroized

1. The MFCC feature matrix is reduced to the first 1000 and the last 1000 measurement points, which results in a 13 by 2000 matrix

System Architecture



Intermediate Results

	Epochs	Accuracy (%)	Trainable Embeddings	Learning Rate	Activation func.	Loss func.
Lexical	15	72.2414	False	0.05	TanH	Hinge-Loss
	15	72.4881	False	0.05	TanH	Hinge-Loss
	15	77.8655	True	0.05	TanH	Hinge-Loss
Acolex	15	69.824	False	0.01	Sigmoid	Hinge-Loss
	dummy	77.8655	True	0.05	TanH	Hinge-Loss
	dummy	77.8655	True	0.05	TanH	Hinge-Loss

Potential Future Work

- What we plan next:
  - Varying MFCC feature size
  - Including word2vec features for the test and development set
  - Implementation of an additional embedding layer between CNN output and output layer

References

[1] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1459–1462, New York, NY, USA, 2010. ACM.

[2] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.