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”)
- Two kinds of features:
  - 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 into training, development and test data
- The acoustic dataset includes a recording for every utterance

		Channel				
		opinion	question	backchannel	statement	Sum
Dataset	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 embedding layer we use the pre-trained Google word2vector model
- Contains 3 million words, representing each word as a 300-dimensional vector

Data Preprocessing

Embedding Matrix

1. All words of the corpus are inserted into an embedding matrix
2. Each word is represented by its corresponding vector from *word2vec*
3. If a word is not contained in *word2vec* it gets assigned a random vector
4. Unkown word and no word vectors added

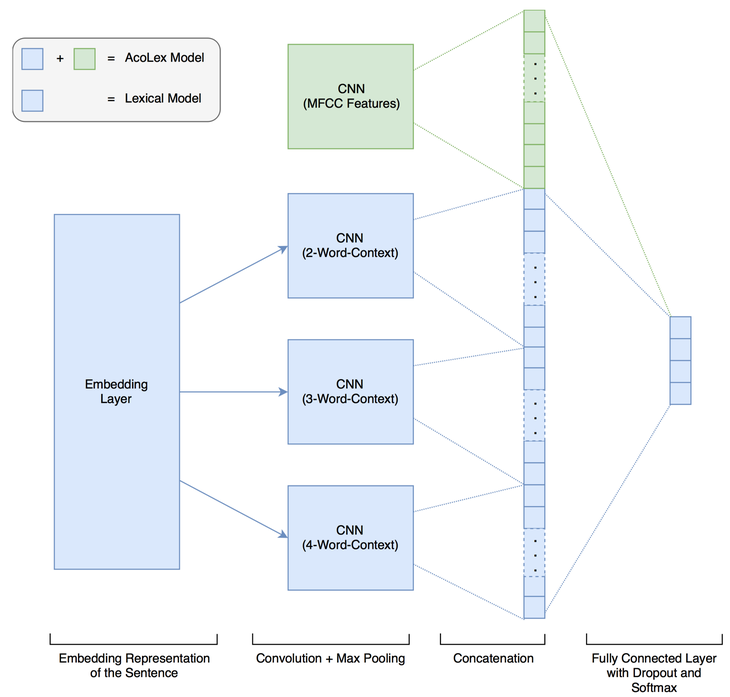
Lexical Features

1. Each sentence is converted into a sequence of indices
2. Each index is the corresponding line in the embedding matrix for one word
3. The maximum sentence length is set to 100 words

Acoustic Features

1. The MFCC features of the first 10 seconds are used
2. The MFCC feature of the last 10 seconds are used
3. If audio file is shorter than 20 seconds the missing MFCC features are zeroized

System Architecture



Intermediate Results

	Epochs	Dropout	Accuracy (%)	Trainable Embeddings	Learning Rate	Activation function	Loss function
Lexical	15	0.5	78.42	True	0.05	TanH	Hinge-Loss
	15	0.5	70.89	False	0.05	TanH	Hinge-Loss
	15	0.5	77.73	True	0.05	Relu	Hinge-Loss
	15	0.5	60.50	False	0.05	Relu	Hinge-Loss
	15	0.5	76.86	True	0.05	Sigmoid	Hinge-Loss
	15	0.5	67.85	False	0.05	Sigmoid	Hinge-Loss
	15	0.5	78.56	True	0.01	TanH	Cross Entropy
	15	0.5	78.66	True	0.05	TanH	Cross Entropy
	15	0.5	69.62	False	0.05	TanH	Cross Entropy
	15	0.5	77.09	True	0.01	Relu	Cross Entropy
	15	0.5	78.10	True	0.05	Relu	Cross Entropy
	15	0.5	70.24	False	0.05	Relu	Cross Entropy
	15	0.5	77.14	True	0.01	Sigmoid	Cross Entropy
	15	0.5	78.98	True	0.05	Sigmoid	Cross Entropy
	15	0.5	66.11	False	0.05	Sigmoid	Cross Entropy
Acolex	15	0.5	77.90	True	0.01	TanH	Hinge-Loss
	15	0.5	73.18	False	0.01	TanH	Hinge-Loss
	15	0.5	74.58	True	0.01	Relu	Hinge-Loss
	15	0.5	62.26	False	0.01	Relu	Hinge-Loss
	15	0.5	73.34	True	0.01	Sigmoid	Hinge-Loss
	15	0.5	61.67	False	0.01	Sigmoid	Hinge-Loss

Potential Future Work

- Varying MFCC feature amount
- Using smoothed training data to better balance the classes
- Using stop word filtering
- Insertion of an additional fully connected layer between the CNN output and the softmax 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.