Lexico-acoustic Neural-based Models for Dialog Act Classification

Daniel Ortega, Ngoc Thang Vu
Institute for Natural Language Processing - University of Stuttgart, Germany
{daniel.ortega, thang.vu}@ims.uni-stuttgart.de

Motivation

Explore the role and usefulness of the acoustic information for dialog act (DA) classification in combination with lexical features by means of neural-based models

Utterance	Dialog Act (DA)
A: *Are you a musician yourself?*	Yes-no-question
B: *Uh, well, I sing.*	Affirmative non-yes answer
A: *Uh-huh.*	Acknowledge (Backchannel)
B: *I don't play an instrument.*	Statement-non-opinion

Models

LEXICAL MODEL (LM)

The LM processes the transcripts of the current utterance and its context using convolutional neural networks (CNNs) and the context learning method RNN-Output-Attention

ACOUSTIC MODEL (AM)

The AM is a CNN-based model to process acoustic features – 13 Mel-frequency cepstral coefficients (MFCC) per frame

LEXICO-ACOUSTIC MODEL (Lex-Ac)

The Lex-Ac model is a bi-CNN that employs lexical and acoustic features and concatenates the outputs of the LM and AM models

Convolutional Neural Network

- The input matrices represent the utterances with lexical or acoustic features
- The CNN performs a discrete convolution with 2D filters f

$$(w*f)(x,y) = \sum_{i=1}^{d} \sum_{j=-|f|/2}^{|f|/2} w(i,j)\cdot f(x-i, y-j)$$

RNN-Output-Attention (ROA)

- ROA is a context learning method that models the relation between the current utterance and its context
- ROA consists of an LSTM followed by a weighted sum of the hidden states h using global attention

Global Attention

For each hidden state $h(t-i)$ at time step $t-i$, the attention weight α_i is:

$$\alpha_i = \frac{exp(f(h(t-i)))}{\sum_j^m exp(f(h(t-j))}$$

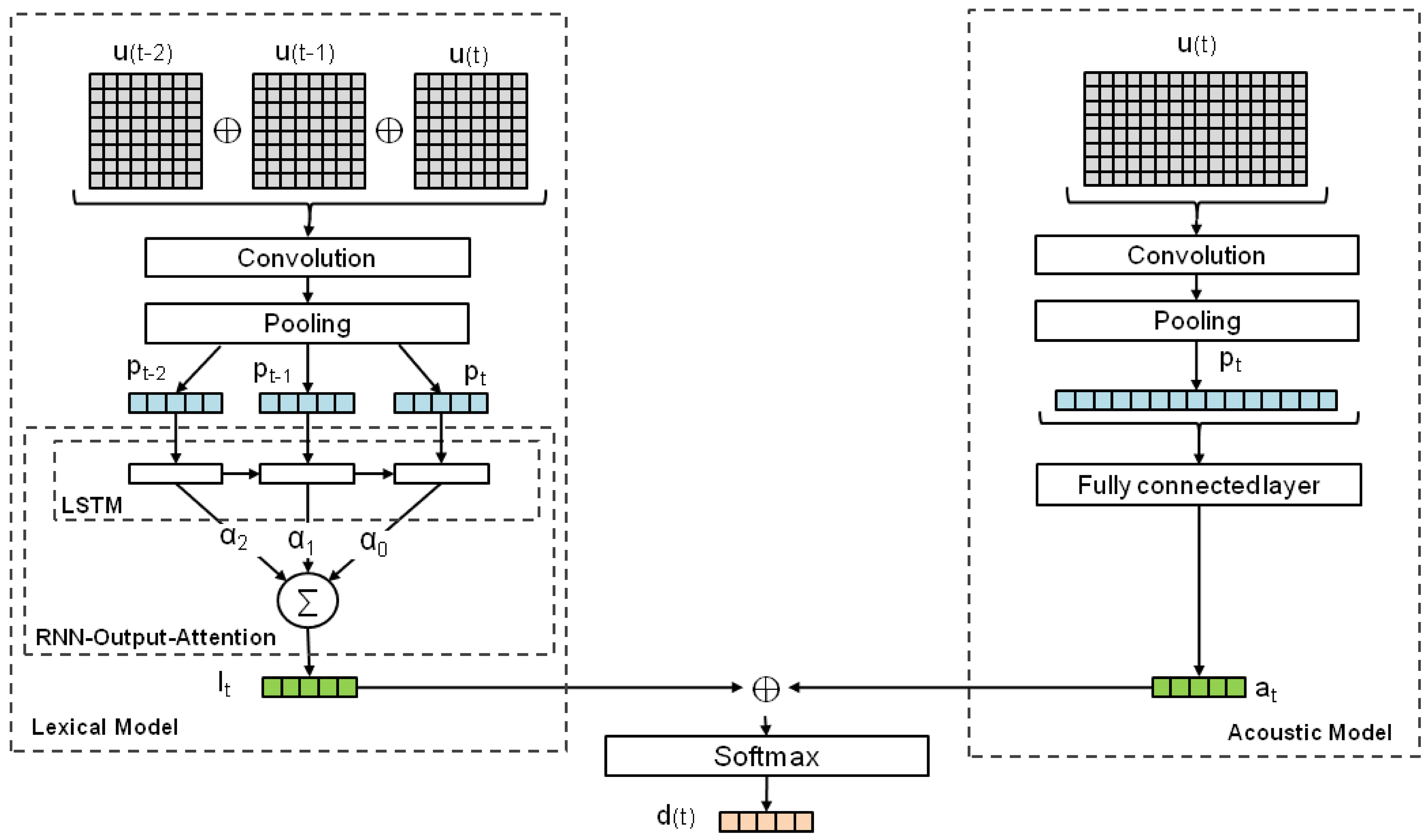
where f is the scoring function, a linear function of the input $h(t-i)$

$$f(h(t-i)) = W^T h(t-i)$$

where W is a trainable parameter. The output l_t is the weighted sum of the hidden sequence

$$l_t = \sum_i \alpha_i h(t-i)$$

Lexico-acoustic Model Architecture



Experimental Setup

Datasets:

MRDA: ICSI Meeting Recorder DA Corpus

SwDA: NXT-format Switchboard DA Corpus

Dataset	C	|V|	Train	Val	Test
MRDA	5	12k	78k	16k	15k
SwDA	42	16k	98k	8.5k	2.5k

C: Number of classes |**V**|: Vocabulary size

Hyperparameter	LM	AM
Filter width	3, 4, 5	5
Feature maps	100	100
Dropout rate	0.5	0.5
Activation function	ReLU	ReLU
Pooling size	utterance-wise	(18,1)
Word embeddings	word2vec	—
MFCC features	—	13
Mini-batch size	50 (MRDA) – 150 (SwDA)	

Results

Accuracy per Model

Model	MRDA	SwDA
Lexical	84.1	73.6
Acoustic	67.8	50.9
Lex-Ac	**84.7**	**75.1**

Single-word Utterances

DA-Right	Lexical	Lex-Ac
Statement	0.45	**0.52**
Backchannel	**0.67**	0.65

F_1 score for utterances *Right* on MRDA

DA-Yeah	Lexical	Lex-Ac
Statement	0.46	**0.57**
Backchannel	0.72	**0.74**

F_1 score for utterances *Yeah* on MRDA

Comparison with Other Works

Model	MRDA	SwDA
Lex-Ac model	**84.7**	**75.1**
NCRL	84.3	73.8
CNN-FF	84.6	73.1
HBM	81.3	—
HCNN	—	73.9
HMM	—	71.0
Majority class	59.1	34.7

NCRL: Neural context representation, CNN-FF: Contextual information on CNNs, HBM: Hidden backoff model, HCNN: Hierarchical CNN, HMM: Hidden Markov model

Effect of Removing the Question Mark

Question	Lexical	Lex-Ac
With ?	**97.7**	96.1
Without ?	46.6	**50.2**

Accuracy (%) for DA *Question* on MRDA

Conclusions

- We proposed an approach to incorporate lexical and acoustic features in a neural model for DA classification
- Our experiments reveal that adding acoustic information to the model improves the overall accuracy and specially helps when:
  - The data for a particular DA is large enough, lexical information is limited and strong lexical cues are not present