Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart

Dialog-act classification using Convolutional Neural Networks

Daniel Ortega

Agenda

- Background
- *Models for Dialog-act classification*
  - *Lexical model*
  - *Acoustic model*
  - *Lexico-acoustic model*
- *Corpora*
- *Results*
- *Conclusion*
- *Future Work*

Background

- **Dialog-act** (DA):
  - Each **utterance** in a dialog has a **performative function** in communication.
  - Dialog-act is an **act of communication** that expresses certain **attitude**:
    - statement → belief, request → desire, apology → regret.
  - A dialog-act succeeds if the audience identifies the **speaker's intention**.

Kent Bach(2000)

Background

- **Dialog-act** (DA):
  - Each **utterance** in a dialog has a **performative function** in communication.
  - Dialog-act is an **act of communication** that expresses certain **attitude**:
    - statement → belief, request → desire, apology → regret.
  - A dialog-act succeeds if the audience identifies the **speaker's intention**.

Kent Bach(2000)

Speaker	Dialog Act	Utterance
A	Wh-Question	What kind do you have now?
B	Statement	Uh, we have a, a Mazda nine twenty nine and a Ford Crown Victoria and a little two seater CRX
A	Acknowledge	Oh, okay.

A fragment of a labeled switchboard conversation

Background

How to approach the task?:

- Lexical approach
  - Traditional approach, it employs the utterance transcription (word sequence)
  - Kim (2014) proposed a sentence classification model: a simple but strong one-layer convolutional neural network using pre-trained word vectors.

Background

How to approach the task?:

- Lexical approach
  - Traditional approach, it employs the utterance transcription (word sequence)
  - Kim (2014) proposed a sentence classification model: a simple but strong one-layer convolutional neural network using pre-trained word vectors.

- Acoustic approach
  - Shriberg et al. (2000) was one of the first works that explored the prosody as a potential knowledge source for dialog-act classification.
  - The DAs can be ambiguous if only lexical information is considered

    Example: *This is your car (?)* → Statement or declarative questions
  - In dialog systems, automatic speech recognizers generate noisy transcriptions, the DA classifier must deal with them.

Background

How to approach the task?:

- Lexical approach
  - Traditional approach, it employs the utterance transcription (word sequence)
  - Kim (2014) proposed a sentence classification model: a simple but strong one-layer convolutional neural network using pre-trained word vectors.
- Acoustic approach
  - Shriberg et al. (2000) was one of the first works that explored the prosody as a potential knowledge source for dialog-act classification.
  - The DAs can be ambiguous if only lexical information is considered

    Example: *This is your car (?)* → Statement or declarative questions
  - In dialog systems, automatic speech recognizers generate noisy transcriptions, the DA classifier must deal with them.
- Lexico-acoustic approach

Background

- **Convolutional Neural Networks** (CNN):
  - CNNs are several layers of convolutions with nonlinear activation functions.
  - Each layer applies different filters and combines their results to obtain high-level features.
  - The last layer is then a classifier that uses these high-level features.
  - Grid-like input format

Background

- **Convolutional Neural Networks** (CNN):

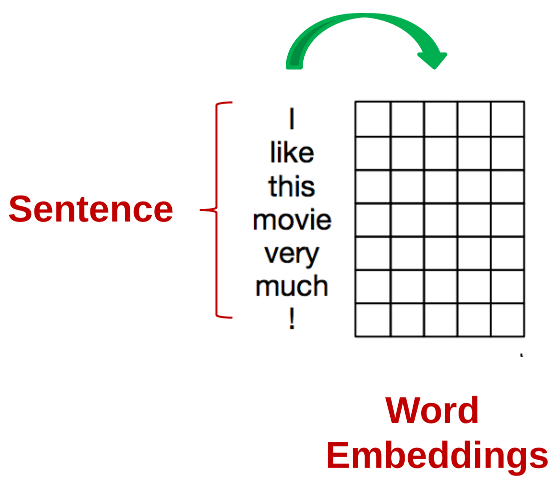


Image source: Zhang. Y., Wallace, B. (2015)

Background

- **Convolutional Neural Networks** (CNN):

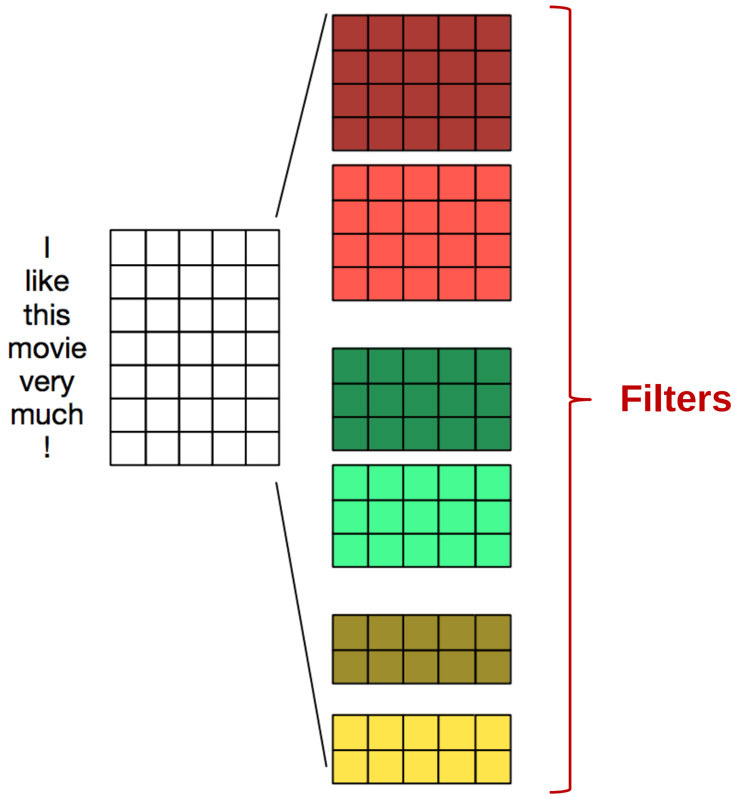


Image source: Zhang. Y., Wallace, B. (2015)

Background

- **Convolutional Neural Networks** (CNN):

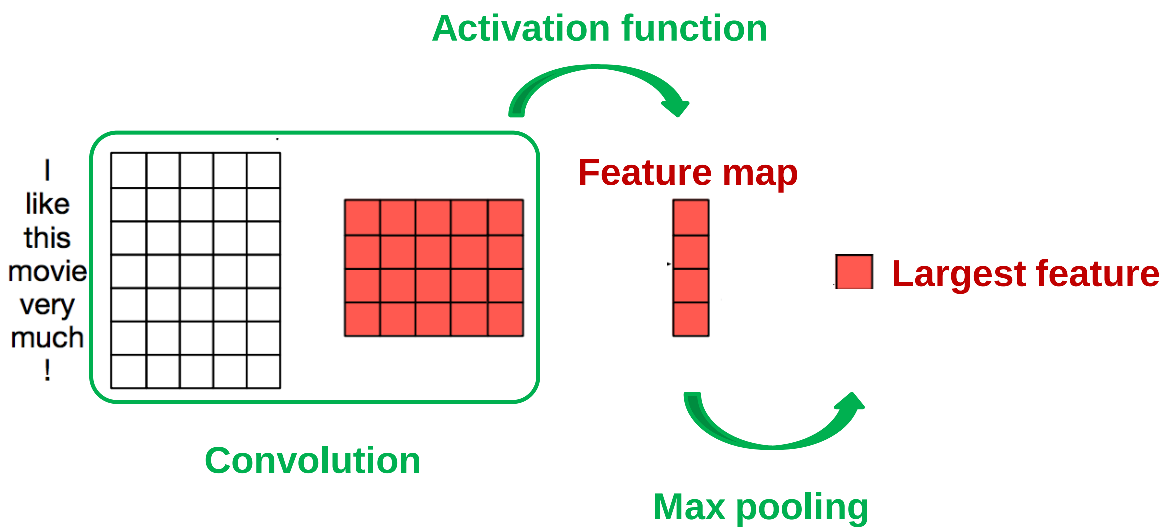


Image source: Zhang. Y., Wallace, B. (2015)

Background

- **Convolutional Neural Networks** (CNN):

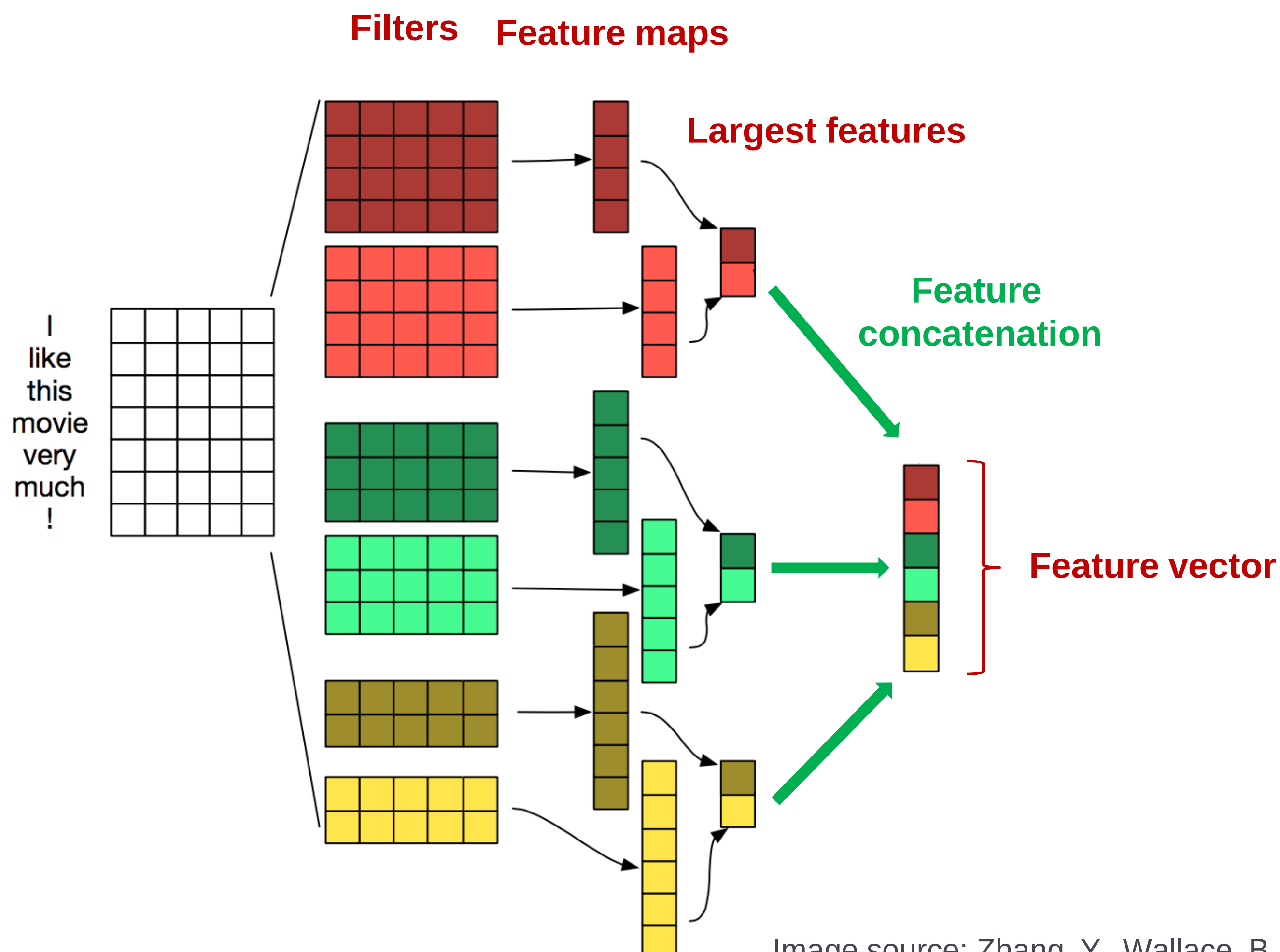


Image source: Zhang. Y., Wallace, B. (2015)

Background

- **Convolutional Neural Networks** (CNN):

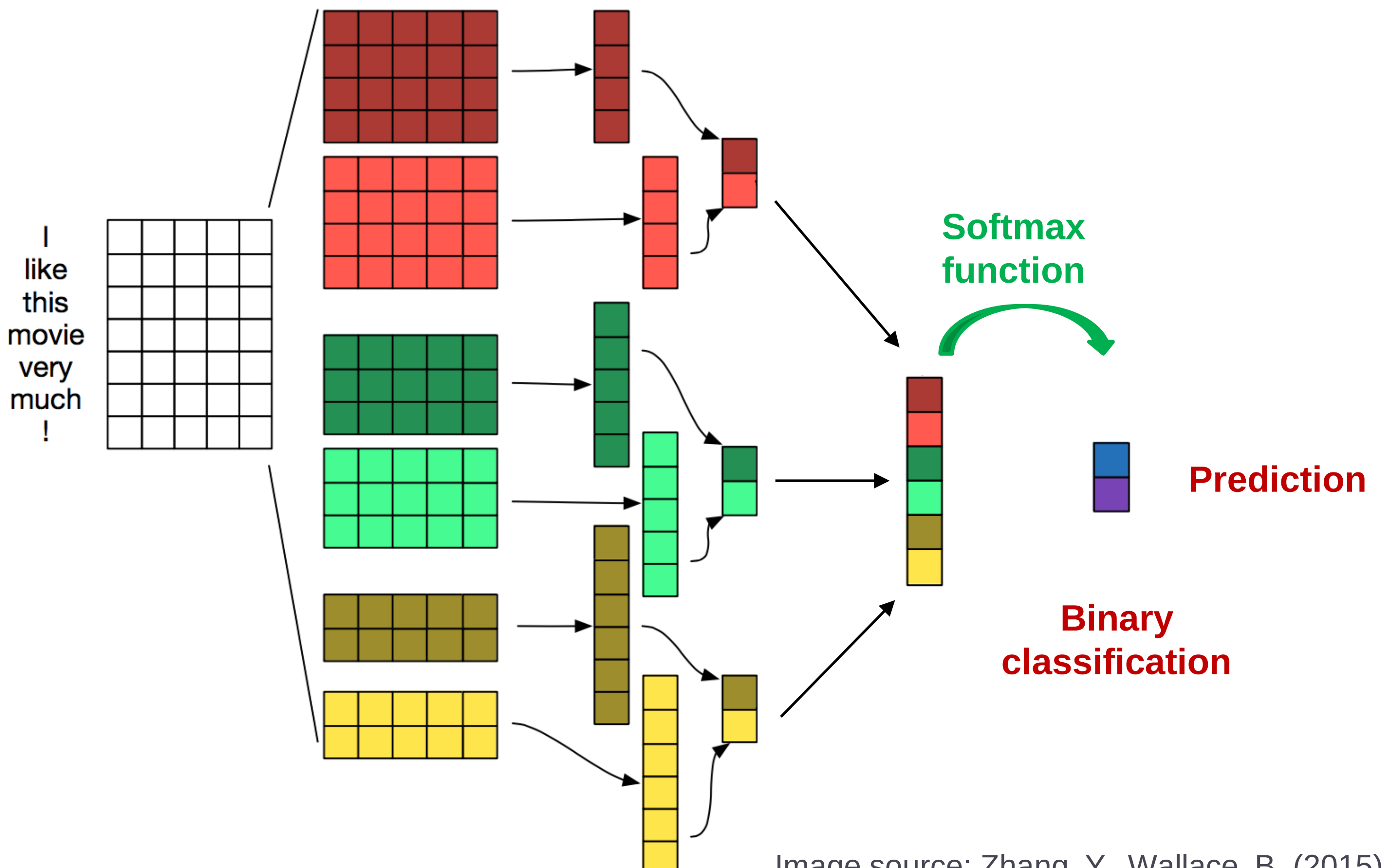
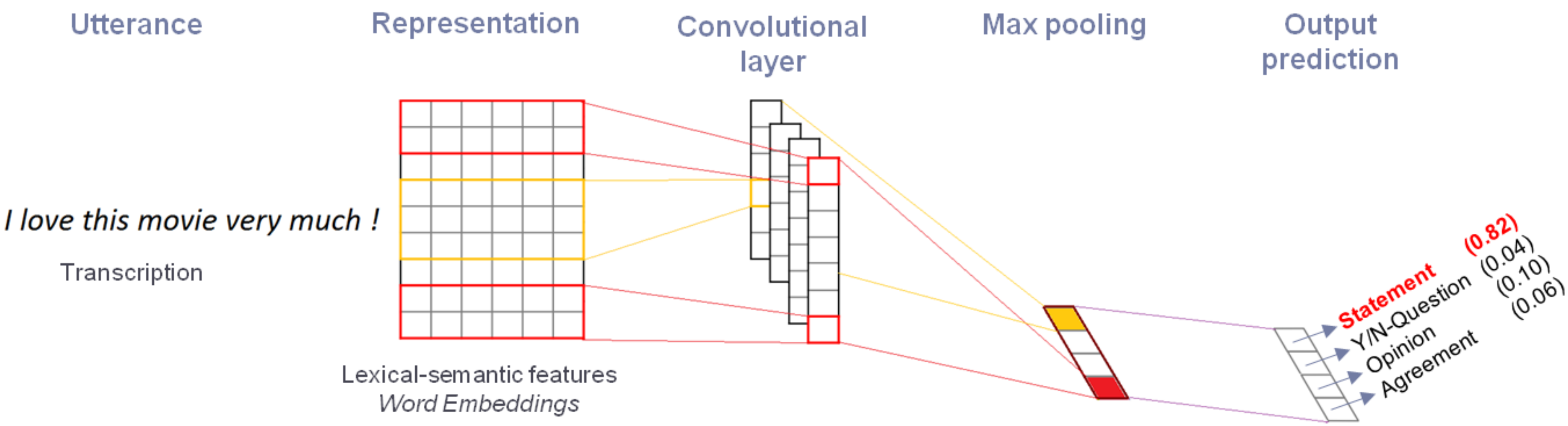


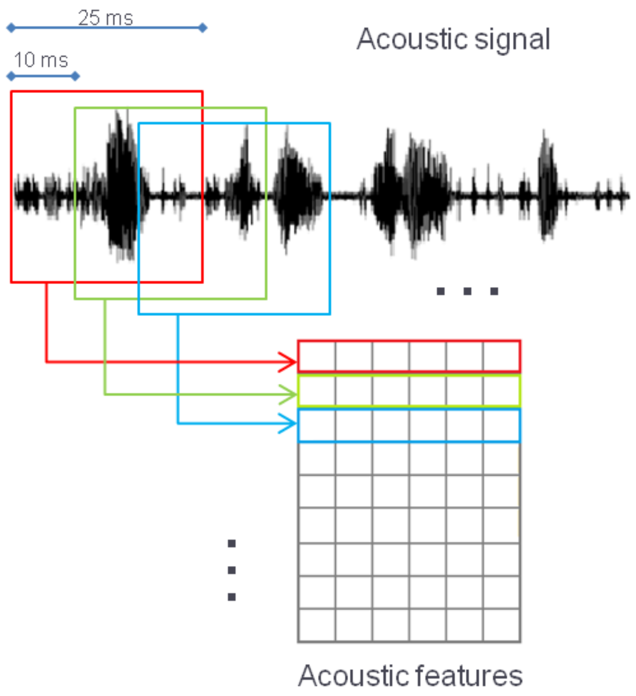
Image source: Zhang. Y., Wallace, B. (2015)

Lexical Model



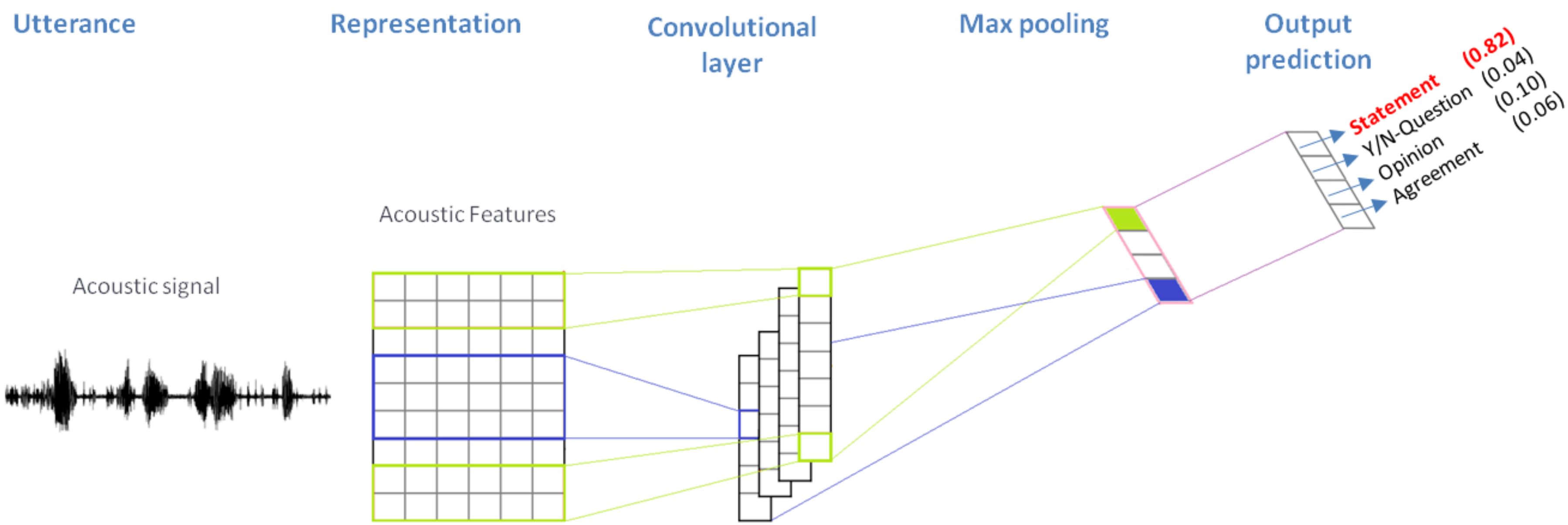
Acoustic Model

- Acoustic feature extraction



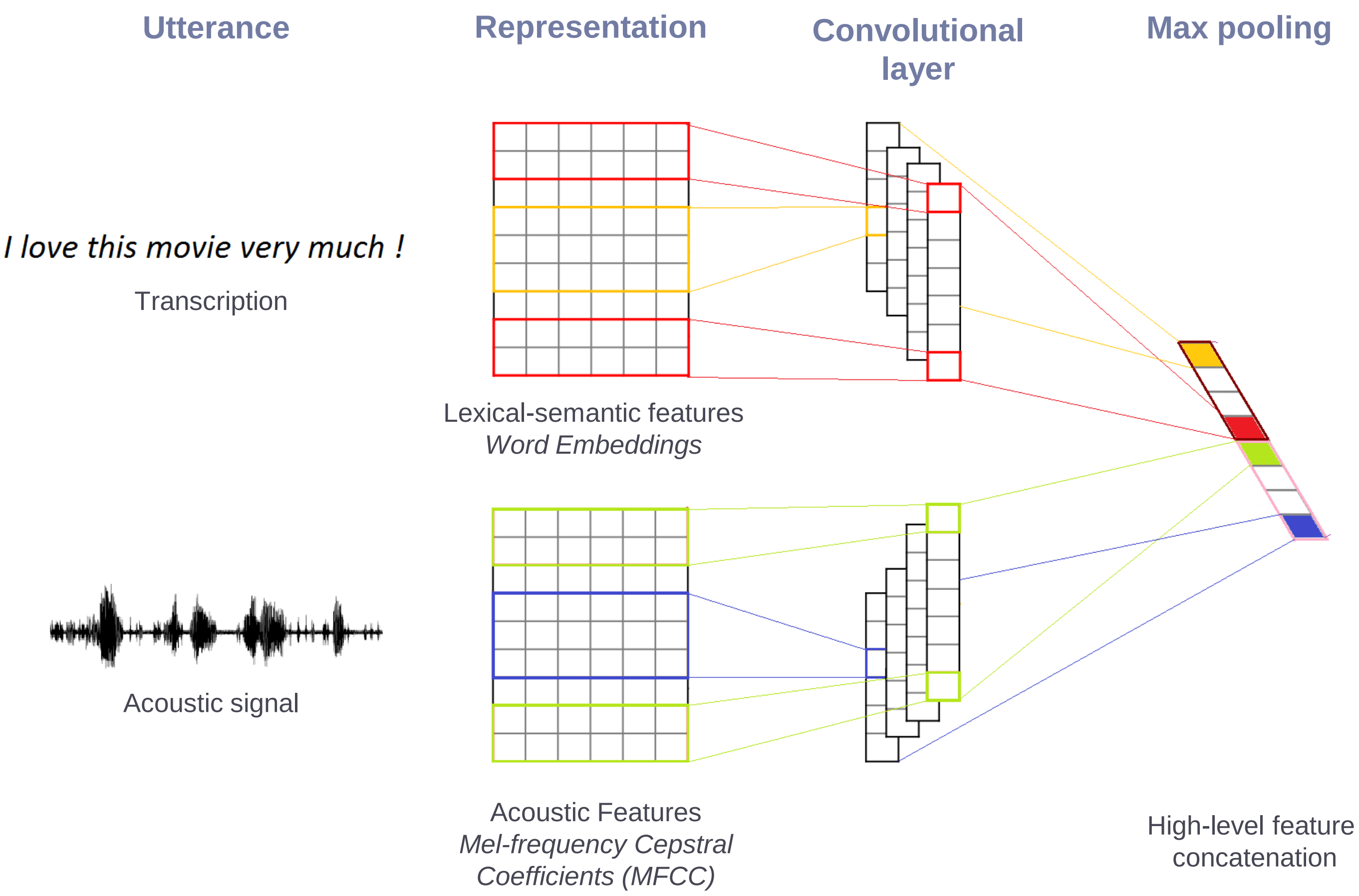
- openSMILE feature sets:
  - Prosodic features: F0, voicing probability and loudness contours
  - LogMel Spectrum
  - Mel-Frequency-Cepstral Coefficients (MFCC)

Acoustic Model



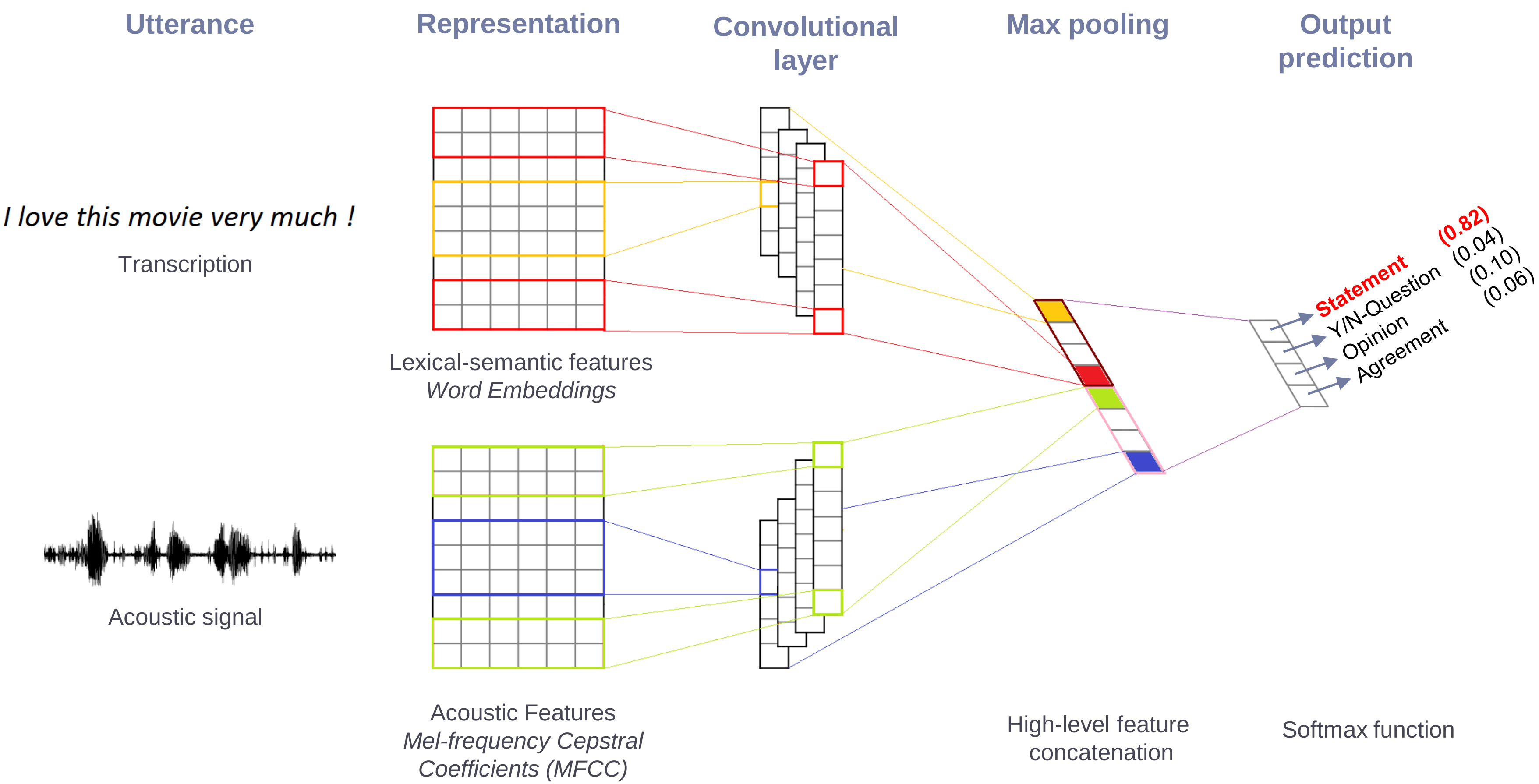
Proposed Model

- **Dialog-act Classifier – Bi-Convolutional Neural Network**



Proposed Model

- **Dialog-act Classifier – Bi-Convolutional Neural Network**



Corpora

Corpus	Training	Test	Classes
ATIS	~5,000	~900	17
Switchboard	~98,000	~10,000	42
ICSI	~99,000	~10,000	5

Results: Lexical Model

Corpus	Classes	Time elapsed per epoch on avg.	Accuracy (%)
ATIS	17	$\sim$ 12 s	94.68
Switchboard	42	$\sim$ 200 s	71.57
ICSI	5	$\sim$ 94s	84.45

Table 6.2: Accuracy per corpus on lexical model

Results: Acoustic Model per Feature Set

Epochs	Accuracy (%) per feature set		
	Prosodic	Log-Mel	MFCC
25	72.40	73.19	**74.88**
50	72.51	**73.71**	73.64
100	72.62	74.09	**75.67**
200	72.74	76.01	**76.24**

ATIS

Epochs	Accuracy (%) per feature set		
	Prosodic	Log Mel	MFCC
5	52.48	52.46	**53.28**
15	52.36	52.35	**53.45**
25	52.46	52.69	**53.55**
50	52.41	52.78	**53.41**

Switchboard

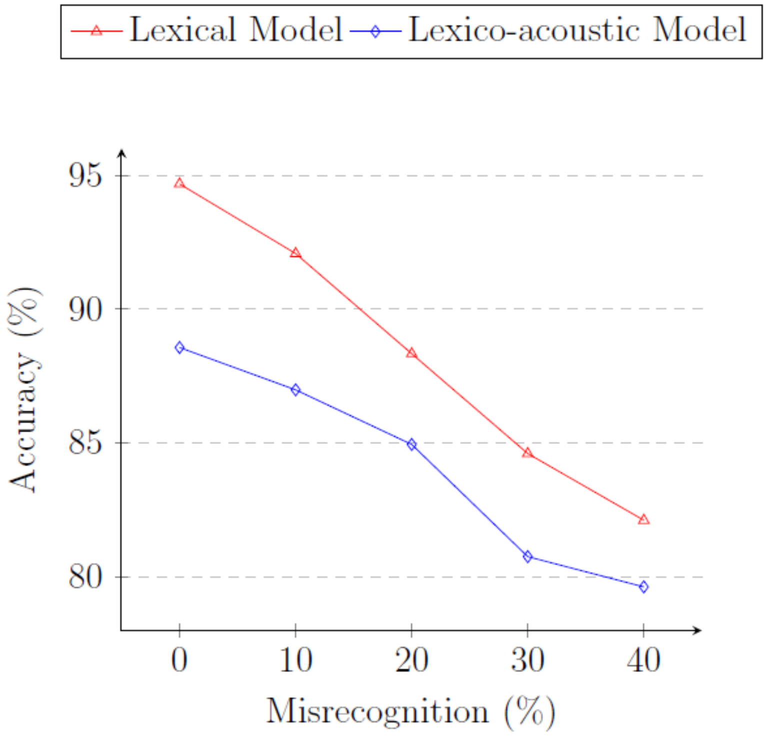
Results on ATIS

Accuracy(%)		
Lexical Model	Acoustic Model	Lexico-acoustic Model
94.68	74.88	88.57

Table 6.7: Accuracy on ATIS per model

Lexico-acoustic Model – ASR emulation

Misrecognition (%)	Accuracy (%) per model	
	Lexical Model	Lexico-acoustic Model
0	94.68	88.57
10	92.08	86.99
20	88.34	84.95
30	84.61	80.76
40	82.12	79.63



ATIS

Results on Switchboard

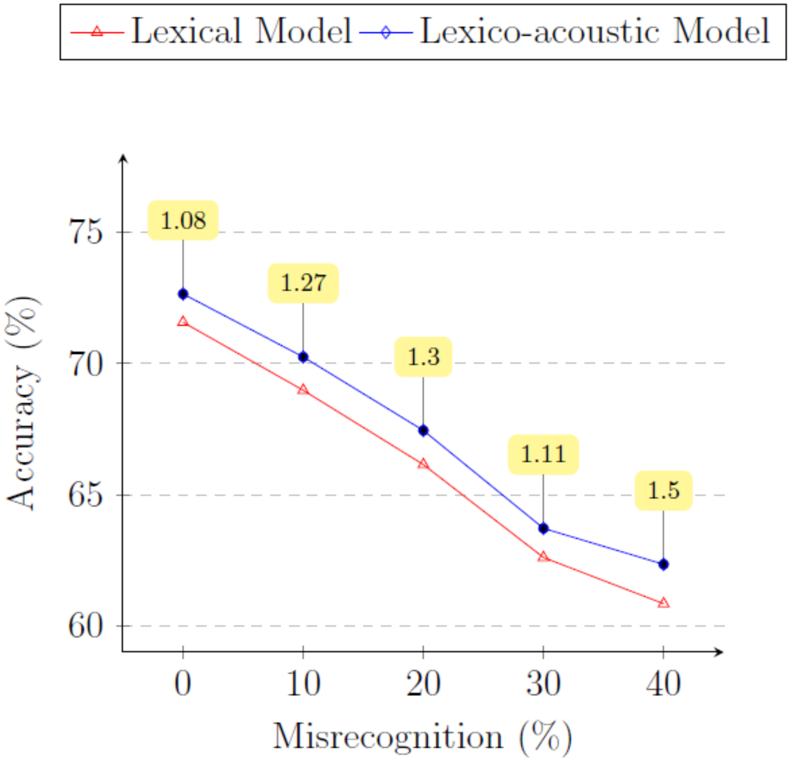
Accuracy(%)		
Lexical Model	Acoustic Model	Lexico-acoustic Model
71.57	53.55	**72.65**

Table 6.10: Accuracy on Switchboard per model

Lexico-acoustic Model - Simulated ASR

Misrecognition (%)	Accuracy (%) per model		Improvement
	Lexical Model	Lexico-acoustic Model	
0	71.57	72.65	1.08
10	68.98	70.25	1.27
20	66.15	67.45	1.3
30	62.61	63.72	1.11
40	60.85	62.35	1.5

Switchboard



The acoustic CNN offsets slightly the recognition error

The larger the misrecognition is, the larger improvement the lexico-acoustic model yields.

Lexico-acoustic Model

Corpus	Classes	CNN			
		Majority Class	Lexical	Acoustic	Bi-CNN
ATIS	17	72.60	**94.68**	74.88	88.57
Switchboard	42	36.00	71.57	53.55	**72.65**
ICSI	5	58.90	**84.45**	—	—

Table 6.12: Accuracy results per model on ATIS, SWBD and ICSI.

Conclusion

- Acoustic Model:
  - MFCC features are more suitable for dialog-act classification.
  - ATIS: accuracy is not significant 3.64% over the majority class.
  - SWBD: accuracy is 17.55% over the majority class.
- Lexico-acoustic Model
  - ATIS: acoustic features worsened the accuracy in 6.5%.
  - SWBD: acoustic features yielded an improvement of 1.08%
  - The acoustic features helped keep the accuracy higher in at least 1.1% regardless of the amount of error (ASR emulation).

Conclusion

- Why contradicting results?
  - In ATIS the utterances are only information requests and the classes are related only to the lexical content,
  - In SWBD the classes are also related to the prosodic content.
  - The utterances in SWBD differ more acoustically themselves and contain phonetic cues that are strongly related to some of the acts.
  - The success of the model depends on the corpus, this is the relation between the prosody in the utterances and the classes.

Future Work

- Combine acoustic feature sets in order to find if there is a more appropiate set
- Train the lexico-acoustic model on ICSI that is similar to SWBD
- Explore Attention Mechanisms on Neural Networks for sentence modeling yielding promising results, in order to highlight words or phrases that are useful for the dialog-act classification.
- Encode sentence context. a Wh-Question is more likely to be followed by a Statement than by another Wh-Question.

Questions…

… Thanks

References

- Kent Bach, Routledge (Firm). **Concise Routledge Encyclopedia of Philosophy**. Psychology Press. Pages 855-856, 2000.
- Yoon Kim. **Convolutional Neural Networks for Sentence Classification.** 2014.
- Elizabeth Shriberg et al. **Can prosody aid the automatic classification of dialog acts in conversational speech?** CoRR, cs.CL/0006024, 2000.
- Ye Zhang and Byron C. Wallace**. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification**. CoRR, abs/1510.03820, 2015.