Modeling Long-Range Context for Concurrent Dialogue Acts Recognition

Yue Yu* [1] Siyao Peng* [2] Grace Hui Yang [1]

[1]Department of Computer Science [2]Department of Linguistics
Georgetown University

CIKM 2019

The Need for a Sequence Model

Table 1: Predictions on a sample dialogue with long-range dependencies.[6]

#	Utterance	toks	Reference	Prev. SOTA [6]	Our model
1	*U1* How can I download Skype for Windows 8.1 ...	32	OriginalQuestion	OriginalQuestion	OriginalQuestion
2	*A1* Hi...if you are using a phone running Windows 8.1 and lower...no longer supported...but if you are using a Windows computer, you can still download...	77	Greetings PotentialAnswer	Greetings PotentialAnswer	Greetings PotentialAnswer
	... Utterances 3 (27 toks) & 4 (70 toks) ...				
5	*U1* Hi, I am using **Surface tablet 8.1 windows** and have tried many times to install the app. But it comes up - This app can't run on this pc please use the store app. But Skype does not appear on here.	46	Greetings **FurtherDetails** **FollowupQuestion**	Greetings **OriginalQuestion**	Greetings **FurtherDetails** PotentialAnswer **RepeatQuestion**
	... Utterance 6 (16 toks) ...				

Our model sees long-range context and performs better on Utterance 5:

- the 5-th utterance should not be an OriginalQuestion;
- *Surface tablet 8.1 windows* provides FurtherDetails;
- FollowupQuestion partially matches RepeatedQuestion.

Highlights

Task

Concurrent Dialogue Acts (CDA) recognition: the task to handle long utterances and concurrent dialogue acts.

Model

Convolutional Recurrent Neural Network (CRNN): Our sequence model that imposes fewer restrictions on the structure of DAs and captures textual features from a wider context.

Dataset

MSDialog-Intent: A tech forum dataset from Microsoft Dialogue Intent Corpus [6] consisting of 10,020 utterances with 72 tokens and 1.83 DAs per utterance.

Convolutional Recurrent Neural Network (CRNN)

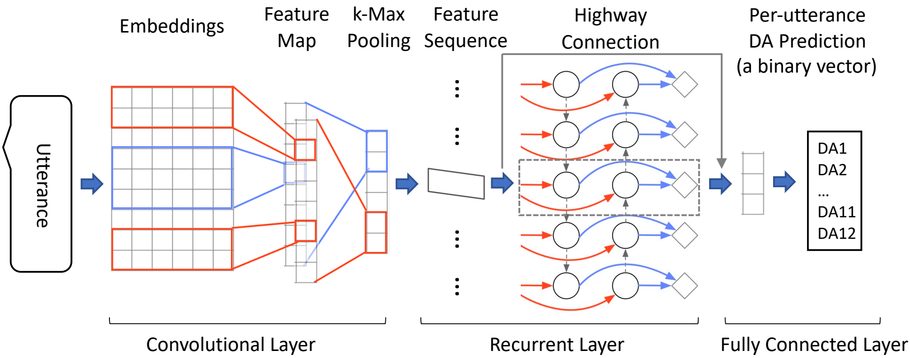

Figure 1: Our proposed CRNN model architecture.

CRNN has been applied to multi-label sequence classifications, including multiple sound event detection [1] and multi-label music tagging [3].

CRNN – Convolutional Layer

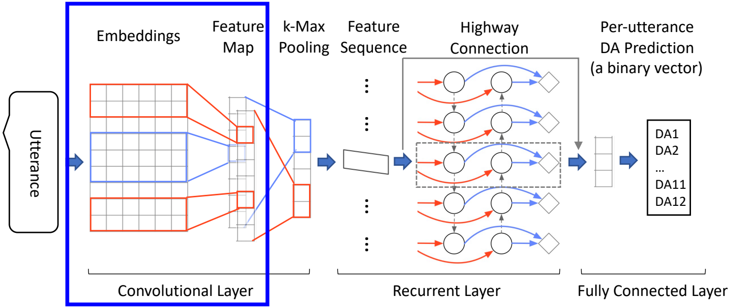

Figure 2: Our proposed CRNN model architecture.

The basic CNN [4] module slides through the embedding matrix of an utterance and and generates a feature map **k**, capturing semantic features in differently ordered n-grams.

$$\mathbf{k} = [k_1, k_2, ..., k_{n-d+1}] \quad (1)$$

CRNN – Dynamic *k*-Max Pooling

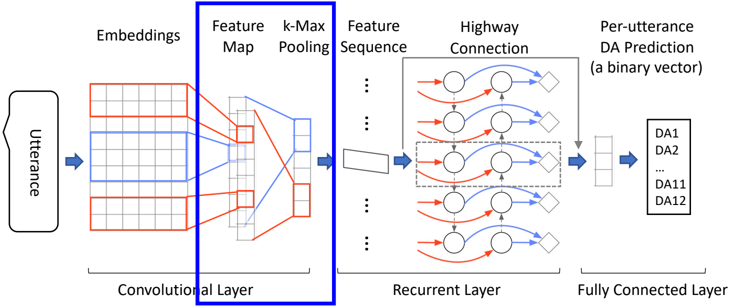

Figure 3: Our proposed CRNN model architecture.

We use Dynamic *k*-Max Pooling [5] to pool the most powerful features from p sub-sequences of an utterance with m words which accommodates variable utterance length.

$$p(\mathbf{k}) = \left[\max\left\{\mathbf{k}_{1:\lfloor \frac{m}{p} \rfloor}\right\}, \ldots, \max\left\{\mathbf{k}_{\lfloor m-\frac{m}{p}+1 \rfloor : m}\right\}\right] \quad (2)$$

CRNN – Recurrent Layer

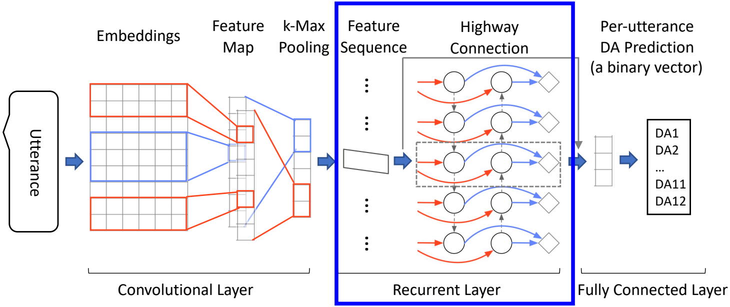

Figure 4: Our proposed CRNN model architecture.

Bidirectional RNNs, both LSTM [8] and GRU [2], are applied to gather features from a wider context in the Feature Sequence for recognizing the DAs in the target utterance, u_t.

CRNN – Highway Connection

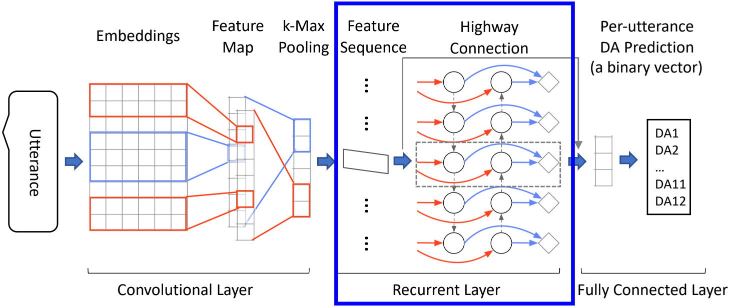

Figure 5: Our proposed CRNN model architecture.

We add Highway Connections [7] between the Convolutional Layer and the Fully Connected Layer so that the information about the target utterance, u_t, can flow across the Recurrent Layer without attenuation.

Experiments

Two Baselines:

- **CNN-Kim[4]**.
- **CNN-CR[6]:** The SOTA CNN with fixed context window.

Our three CRNN experiments with incremental improvements:

- **CRNN (v_1):** Our base model with Binary Cross Entropy (BCE) loss and sigmoid activation function.
- **CRNN (v_2):** CRNN (v_1) + highway connections.
- **CRNN (v_3):** CRNN (v_1) + highway connections + dynamic k-max pooling.

Results – Overall Performance

Our CRNN models (v_3 especially) outperform both baselines in terms of:

1. Highest accuracy, recall and F_1 with LSTM; and precision with GRU (Table 2).

Models	**Accuracy**	Precision	Recall	F_1 score
CNN-Kim[4]	0.5785	0.6371	0.6745	0.6553
CNN-CR[6]	0.6354	0.7108	0.6952	0.7029
CRNN (v_1) w/ LSTM	0.6668*	0.7238	0.7297	0.7267
CRNN (v_1) w/ GRU	0.6543*	0.7056	0.7065	0.7061
CRNN (v_2) w/ LSTM	0.6731****	0.7315	0.7315	0.7315
CRNN (v_2) w/ GRU	0.6734**	0.7280	0.7334	0.7307
CRNN (v_3) w/ LSTM	**0.6822********	0.7254	**0.7422**	**0.7337**
CRNN (v_3) w/ GRU	0.6733***	**0.7358**	0.7215	0.7286

Table 2: Performance of CNN-Kim, CNN-CR, & CRNN.

* for $p \leq 0.05$, ** for $p \leq 0.01$, *** for $p \leq 0.001$ and **** for $p \leq 0.0001$.

Results - Better on Multi-DAs

2. Higher accuracy for all reference DA sizes (Table 3).
3. The average number of predicted DAs is closer to the reference (Table 4).

# of ref DAs	% in test	Mean accuracy	
		CRNN (v_3)	CNN-CR
1	36.9	**0.7704****	0.7126
2	42.8	**0.6641*****	0.6232
3	16.7	**0.5596***	0.5177
≥4	3.6	**0.5618**	0.5339

Table 3: Mean accuracy per number of reference DAs.

# of ref DAs	% in test	Avg. num. of pred DAs	
		CRNN (v_3)	CNN-CR
1	36.9	1.44	1.44
2	42.8	**2.02****	1.89
3	16.7	**2.56*****	2.37
≥4	3.6	**2.68**	2.74

Table 4: Average number of predicated DAs per number of reference DAs.

* for $p \leq 0.05$, ** for $p \leq 0.01$, *** for $p \leq 0.001$ and **** for $p \leq 0.0001$.

Results – Better on Longer Dialogues

4. Higher mean accuracy for longer dialogues (Figure 3).

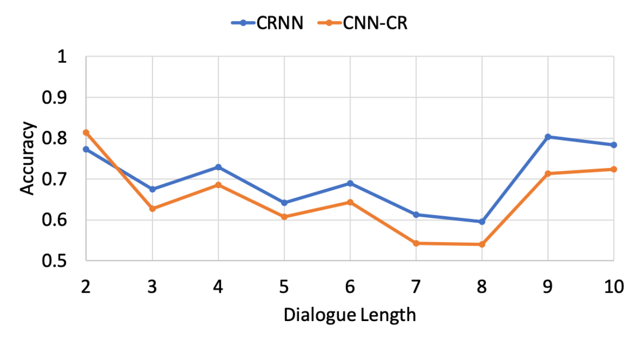

Figure 6: CRNN (v_3) vs. CNN-CR over dialogues of different lengths.

Conclusions

- Our CRNN models achieve the new SOTA for CDA recognition on a tech forum dataset, where the dialogues are packed with complex DA structures and information-rich utterances.
- Our best model significantly outperforms CNN-CR[6] on accuracy by 4.68%; 1.46% on Precision, 4.70% on Recall, and 3.08% on F_1.
- All of our proposed adaptations, i.e. highway connections and dynamic k-max pooling, contribute to the model.

References

[1] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen. 2017. Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (2017).

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[3] K. Choi, G. Fazekas, M. Sandler, and K. Cho. 2017. Convolutional recurrent neural networks for music classification. In *ICASSP '17*. 2392–2396.

[4] Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP'14*. 1746–1751.

[5] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang. 2017. Deep Learning for Extreme Multi-label Text Classification. In *SIGIR '17*. New York, NY, USA, 115–124.

[6] C. Qu, L. Yang, B. Croft, Y. Zhang, J. R. Trippas, and M. Qiu. 2019. User Intent Prediction in Information-seeking Conversations. *CHIIR '19* (2019).

[7] P. Sanders and D. Schultes. 2005. Highway hierarchies hasten exact shortest path queries. In *European Symposium on Algorithms*. Springer, 568–579.

[8] W. Zaremba and I. Sutskever. 2014. Learning to execute. (2014).

Thank you.
Questions?

Paper

TODO-URL