Encoding Word Confusion Networks with Recurrent Neural Networks for Dialog State Tracking

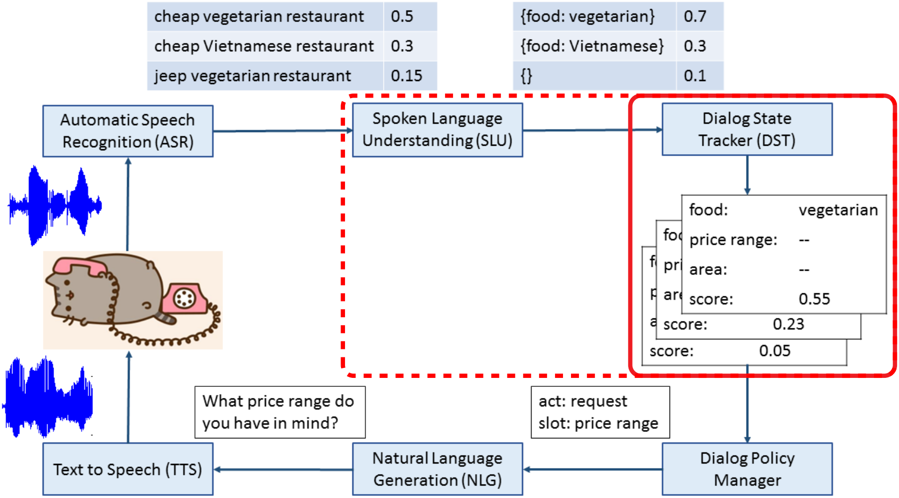
Glorianna Jagfeld, Ngoc Thang Vu

Institute for Natural Language Processing (IMS), University of Stuttgart

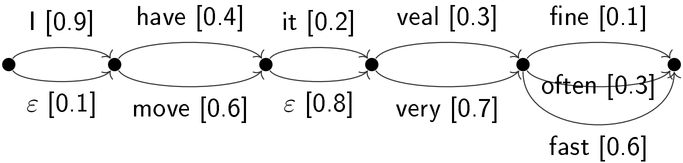
7th of September 2017
SCNLP @ EMNLP 2017 - Short Paper

Motivation

Modular Spoken Dialog System



Word Confusion Network (Cnet)



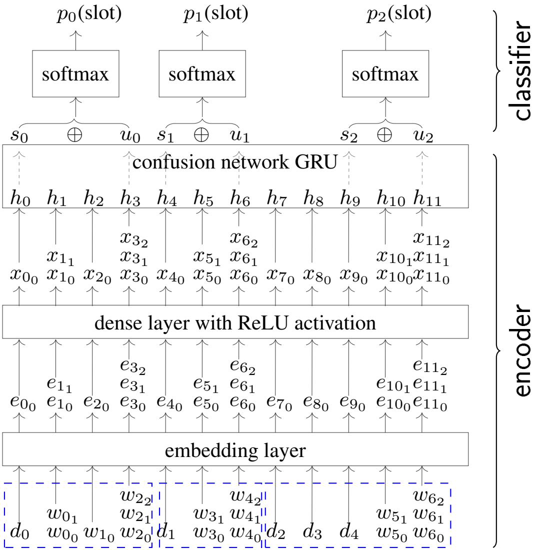
- Richer ASR hypothesis space than n-best list
- More compact data structure than speech lattices
- Every lattice can be converted to a cnet without significant loss of hypotheses (Mangu et al., 2000; ?)

Contributions

- Propose to mitigate damage of ASR errors by using cnets
- First step towards tighter integration of ASR into end-to-end SDSs
- Novel algorithm to encode cnets via recurrent neural network (RNNs) with gated recurrent units (GRUs) (?)
- Show that encoding cnets improves DST performance over ASR 1-best baseline

Model

Model



s_j, u_j: GRU-based cnet encoder outputs at the end of each system and user utterance

$\oplus := W_s s_j + W_u u_j + b$, weighted sum of system and user information

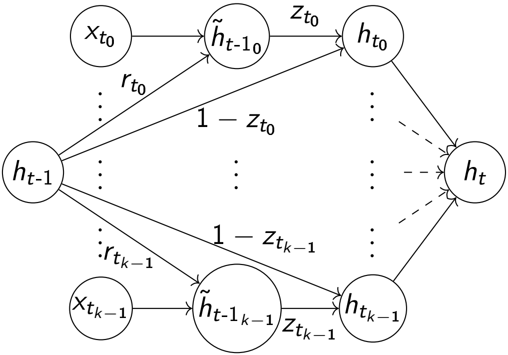
d_t: one-hot vectors of system dialog acts

w_{t_i}: one-hot vectors of word hypotheses in the cnet timesteps of user utterances

Basis: Zilka and Jurcícek (2015)

GRU-based Cnet Encoder

Encoding k alternative hypotheses at timestep t of a cnet:



$$
\begin{aligned}
h_{t_i} &= z_{t_i} \cdot h_{t-1} + (1 - z_{t_i}) \cdot \tilde{h}_{t_i} \\
z_{t_i} &= \sigma(W_z x_{t_i} + U_z h_{t-1} + b_z) \\
\tilde{h}_{t_i} &= \tanh(W_h x_{t_i} \\
&\quad + U_h(r_{t_i} \cdot h_{t-1}) + b_h) \\
r_{t_i} &= \sigma(W_r x_{t_i} + U_r h_{t-1} + b_r) \\
h_t &= f_{\text{pool}}(h_{t_0} \dots h_{t_{k-1}})
\end{aligned}
$$

Based on recent approaches to encode lattices via RNNs (Ladhak et al., 2016; Su et al., 2017)

Choices for the Pooling Function

$$\text{average pooling}: f_{\text{average}} = \frac{\sum_{i=1}^{k} h_{t_i}}{k}$$

$$\text{weighted pooling}: f_{\text{weighted}} = \sum_{i=1}^{k} \text{score}_i \cdot h_{t_i}, \text{ where score}_i \text{ is the}$$

confidence score of cnet hypothesis x_{t_i}

Experiments

Data

- Dataset of the second Dialog State Tracking Challenge (DSTC2) [Henderson et al., 2014]: user interactions with restaurant domain SDS
- 1612 training, 506 development, 1117 test dialogs
- Dialog state: three goals: *area* (7 values), *food* (93 values), *price range* (5 values); 8 requests (e.g. *phone number, address*)
- Train on manual transcripts + cnets, test on cnets
- Represent tokens of system dialog acts, manual transcripts, and 1-best hypothesis as timesteps with single hypothesis
- Cnet preprocessing: 125 hypotheses in average cnet, but average length of best hypothesis is only 4 tokens
  - → remove interjections (uh, oh, ...)
  - → prune hypotheses with low scores

Model Hyper-Parameters

parameter	value
training epochs	20 (requests), 100 (food), 50 (area, price range)
optimizer	Adam (?)
initial learning rate	0.001
training batch size	10 dialogs
λ of l2 regularization	0.001
dropout rate	0.5
embeddings	pretrained 300-dimensional PARAGRAM-SL999 embeddings
# units dense layer	300
# units GRU	100
size of the system and user vector combination matrix	50

Results

Impact of ASR Errors on 1-best Baseline

test data	**goals**	**requests**
train on transcripts + batch ASR (baseline)		
batch ASR	$63.6\,^{66.6}_{58.7}$	$96.8\,^{97.1}_{96.5}$
train on transcripts + live ASR (lower WER)		
live ASR	$63.8\,^{67.0}_{60.2}$	$97.5\,^{97.7}_{97.2}$
transcripts	$78.3\,^{82.4}_{74.3}$	$98.7\,^{99.0}_{98.0}$

DSTC2 test set accuracy of 10 runs with different random seeds in the format average $^{\text{maximum}}_{\text{minimum}}$

→ ASR errors strongly affect DST performance

Results for the Cnet Encoder

method	**goals**	**requests**
1-best baseline	$63.6\ ^{66.6}_{58.7}$	$96.8\ ^{97.1}_{96.5}$
cnet - no pruning		
weighted pooling	$63.7\ ^{65.6}_{61.6}$	$96.7\ ^{97.0}_{96.3}$
cnet - score threshold 0.001		
average pooling	$63.7\ ^{66.4}_{60.0}$	$96.6\ ^{96.8}_{96.0}$
weighted pooling	$\mathbf{65.2}\ ^{68.5}_{59.1}$	$97.0\ ^{97.4}_{96.6}$
cnet - score threshold 0.01		
average pooling	$64.6\ ^{67.9}_{59.7}$	$96.9\ ^{97.2}_{96.5}$
weighted pooling	$64.7\ ^{68.4}_{62.2}$	$\mathbf{97.1}^{\star}\ ^{97.3}_{96.9}$

DSTC2 test set accuracy of 10 runs with different random seeds

*: significantly better than baseline ($p < 0.05$)

Conclusions

Conclusions

- ASR errors pose a major obstacle to accurate DST

→ Leverage richer ASR hypothesis space in cnets

- Novel method to encode cnets by GRU-based RNN: improves DST performance over 1-best baseline

Future Work

- Compare cnet performance against n-best lists
- Explore further ways to leverage the cnet hypothesis scores

Thanks!

gloriannna.jagfeld
thang.vu
@ ims.uni-stuttgart.de

Selected References

Faisal Ladhak, Ankur Gandhe, Markus Dreyer, Lambert Matthias, Ariya Rastrow, and Björn Hoffmeister. 2016. LatticeRNN: Recurrent Neural Networks over Lattices. In *Proceedings of Interspeech*

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4).

Jinsong Su, Zhixing Tan, Deyi Xiong, Rongrong Ji, Xiaodong Shi, and Yang Liu. 2017. Lattice-Based Recurrent Neural Network Encoders for Neural Machine Translation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Lukás Zilka and Filip Jurcícek. 2015. LecTrack: Incremental Dialog State Tracking with Long Short-Term Memory Networks. In *Proceedings of the 18th International Conference on Text, Speech, and Dialogue - Volume 9302*.