SAMSUNG
Best Paper Award 2017

An Attention-based Hybrid Neural Network to Determine Coherent Utterance in Multi-turn Dialogue System

Binger Chen[1]

(Abstract)Multi-turn dialogue is a new and vital technology domain in artificial dialogue system. In existing dialogue system, especially in application area, although machine can interact with human in multi-turn based on context information, it must be in wake-up status or waken up before receiving every utterance through specific command, such as Amazon Echo's waken-up word 'Alexa' or Siri's speaking button. Machine cannot identify the object user is talking to automatically. It must confirm current utterance is coherent before analysis it and make a response. This paper presents a context-aware hybrid neural network to determine whether current dialogue session is aborted. We leverage word embedding model and RNN hidden states to compute the similarity between current user's utterance and every historical utterance in the session. Then we extract matching features in convolutional neural network (CNN) and determine the matching degree between current utterance and the whole dialogue session through attention-based Gated Recurrent Unit (GRU). We consider utterance is related to current dialogue if their matching degree is over a defined threshold, i.e. the conversation continues. We test our model on two open-source multi-turn dialogue corpus, including English and Chinese resources. The Results show that our model can effectively determine the coherent utterance.

1.INTRODUCTION

Dialogue system is developing rapidly, both in academia and in industry. It is becoming quite common in our daily life, especially task-oriented voice assistant which can complete specific task based on colloquial instruction in vertical domain[1]. These dialogue system facilitated human's daily life and even changed lifestyle. Amazon's Alexa, Apple's Siri, Microsoft's Cortana and recently published Samsung's Bixby voice assistant have become the focus of attention[2].

Among the existing works on dialogue system, single-turn dialogue technology is already mature. [3][4][5][6] study the short text single-turn conversation scenario in detail. In real life, multi-turn conversation is the most customary way of interaction. Real dialogue usually won't abort after single-turn, it will continue much longer to convey more information. Therefore, multi-turn dialogue system is considered more and more significant by academia. In addition to extracting information from latest utterance, this system will also take every utterance in current dialogue session, and then construct an overall language model, from which the semantic intent is distilled[7][8][9][10][11].

While most research pay attention to provide correct reply given interaction with user, we study the multi-turn dialogue system in a completely different application domain. Existing experimental or commercial dialogue systems all need a mechanism to ensure the utterance is still in current session. The solutions are either considering the dialogue never ends by default (academia), or leveraging a speaking button (Siri, Bixby), or employing a trigger word (Alexa), which make the smart system rigid and less smart. There's still a huge gap between these systems and real human's free talk. A system which can determine a coherent utterance can dramatically facilitate user's experience. This paper proposes an effective system to analysis the correlation between user's current utterance and dialogue session and dismiss irrelevant ones.

We present a novel end-to-end context-aware hybrid neural network model. The architecture consist of three phases: 1.Matching phase - to compute similarity between input utterance and current dialogue, 2.Similarity analysis phase - to extract similarity features from the matching information, 3.Output phase - to determine a coherent information. We leverage several neural networks to compose a hybrid structure. To be specific: In order to take the whole context into consideration, we match user's utterance with every sentence in current dialogue session. We implement matching process in two levels: First we employ word embedding to compute the similarity of each pair at word level; Second we compute similarity at sentence level through RNN's hidden states[12]. Subsequently, we efficiently extract and aggregate vital matching features from the similarity information achieved in the previous step using CNN's convolutional and pooling mechanism. The next procedure is to integrate the knowledge captured in these matching pairs through GRU and output the matching score of input utterance and the whole dialog session. The gate mechanism of GRU can assist to select non-trivial knowledge for learning. In order to improve computational complexity and pertinence of the algorithm, we introduce the attention mechanism, which recently has been resurgence in computational model[13][14][15][16]. Attention-based GRU can endow weight to the knowledge distilled before[17][18]. We place an attention over all sentences in current dialogue session[19]. It not only decreases the computing cost, but also maintains the vital information. Finally, we define an appropriate threshold to determine a coherent utterance based on matching score.

We experiment our model on a large-scale open source multi-turn dialogue corpus which contains 1 million dialogues and 7 million utterances[20][21]. We randomly sample 4 utterances from the entire corpus as noise for each real dialogue, i.e. our task is to select the correct coherent utterance from 5 candidates. Results show our model can effectively determine whether the dialogue is aborted. In addition, we also test our model on a Chinese multi-turn dialogue corpus which is collected from popular Chinese social media such as Sina Weibo and Douban group forum[22]. Results prove to outperform best baseline model.

In this paper, our main contributes contain: 1.To our best knowledge, we first study the multi-turn dialogue system from a brand-new application aspect C determining coherent utterance. Our results may dramatically improve the user experience of exiting dialogue systems. 2. We propose a novel

[1]Advanced 2 Lab, SRC-Nanjing, Building 4, Chuqiao City, Andemen Street No.57, Nanjing, 210019, China

context-aware attention-based hybrid neural network and verify if on both English and Chinese open source corpus.

REFERENCES

[1] Keizer S et al. Young S, Gai M. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, (2010).

[2] http://www.samsung.com/us/explore/bixby/overview/.

[3] Li H et al. Wang H, Lu Z. A dataset for research on short-text conversations. In *EMNLP*, pages 935–945, (2013).

[4] Li H. Ji Z, Lu Z. An information retrieval approach to short text conversation. *arXiv preprint arXiv*, 1408.6988, (2014).

[5] Li H et al. Wang M, Lu Z. Syntax-based deep matching of short texts. *arXiv preprint arXiv*, 1503.02427, (2015).

[6] Li Z et al. Wu Y, Wu W. Topic augmented neural network for short text conversation. *arXiv preprint arXiv*, 1605.00090, (2016).

[7] Zhou M et al. Wu Y, Wu W. Sequential match network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv*, 1612.01627, (2016).

[8] Tur G et al. Chen Y N, Hakkani-Tr D. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Proceedings of Interspeech*, (2016).

[9] Rochette A et al. Sarikaya R, Xu P. Contextual language understanding for multi-turn language tasks. *U.S. Patent Application*, 14/556,874, (2014).

[10] Wu H. Yan R, Song Y. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.*, pages 55–64, (2016).

[11] Hua W Shiqi Z R Y D Y Xuan L Xiangyang Z, Daxiang D and H T. Multi-view response selection for human-computer conversation. In *EMNLP*, (2016).

[12] Zweig G. Williams J D, Asadi K. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv*, 1702.03274, (2017).

[13] K. Cho D. Bahdanau and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, (2015).

[14] G.Wayne A. Graves and I. Danihelka. Neural turing machines. *arXiv preprint arXiv*, 1410.5401, (2014).

[15] S. Chopra J. Weston and A. Bordesa. Memory networks. In *International Conference on Learning Representations*, (2015).

[16] R. Fergus et al. S. Sukhbaatar, J. Weston. End-to-end memory networks. In *Advances in Neural Information Processing Systems.*, page 24312439, (2015).

[17] Su J. Zhang B, Xiong D. A gru-gated attention model for neural machine translation. *arXiv preprint arXiv*, 1704.08430, (2017).

[18] Cohen W W et al. Dhingra B, Liu H. Gated-attention readers for text comprehension. *arXiv preprint arXiv*, 1606.01549, (2016).

[19] Yin Q et al. Liu T, Cui Y. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. *arXiv preprint arXiv*, 1606.01603, (2016).

[20] Serban I et al. Lowe R, Pow N. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv*, 1506.08909, (2015).

[21] https://github.com/rkadlec/ubuntu-ranking-dataset-creator.

[22] https://github.com/markwunlp/multiturnresponseselection.