The OneConn-MemNN System for Knowledge-Grounded Conversation Modeling

Junyuan Zheng[1], **Surya Kasturi**[1], **Mason Lin**[1], **Xin Chen**[1*], **Onkar Salvi**[1], **Harry Jiannan Wang**[1,2]
[1] OneConnect US Research Institute, New York, NY, 10019
[2] University of Delaware, Newark, DE, 19716

Abstract

End-to-end neural approaches to conversational response generation have been increasingly applied to dialogue systems to help converse with humans naturally and appropriately. Such neural approaches do not rely on manually defined rules and can scale to open domain and free-form datasets. However, most existing end-to-end models can not go beyond chitchat and incorporate entities or factual contents from external knowledge base. In this paper, we present our OneConn-MemNN system participated in the end-to-end conversation modeling task of the Dialog System Technology Challenge 7 (DSTC7), which aims at generating knowledge-grounded conversational responses in a fully data-driven manner. We tailor a memory augmented sequence-to-sequence (SEQ2SEQ) model with attention mechanism to not only represent the human-to-human conversations but also external knowledge sources. We evaluate the response quality using both human evaluation and standard machine translation metrics such as BLEU, NIST, and METEOR. As a result, our system achieves the highest human rating score on appropriateness and the second highest human rating score on informativeness.

1 Introduction

Fully data-driven approaches to conversational response generation have been increasingly applied to dialogue systems to help converse with humans naturally and appropriately. Such neural approaches do not rely on manually defined rules and can easily scale to large-scale conversation corpus. The original end-to-end (E2E) conversation models are inspired by statistical machine translation (Koehn, Och, and Marcu 2003; Och and Ney 2004), including neural machine translation (Kalchbrenner and Blunsom 2013; Cho et al. 2014a; Bahdanau, Cho, and Bengio 2014). Other works in this direction include Long Short-Term Memory (LSTM) models (Vinyals, Fortunato, and Jaitly 2015; Li et al. 2015), the Hierarchical Recurrent Encoder-Decoder (HRED) models (Serban et al. 2016), attention-based models (Yao, Zweig, and Peng 2015; Mei, Bansal, and Walter 2017; Shao et al. 2017), and the Pointer Network model (Gu et al. 2016).

However, existing E2E conversation models often generate bland and deflective responses such as "I'm not understanding" or "I'm not sure", which prevents these systems from effectively engaging the users. Unlike goal-oriented, task-oriented or task-completion dialogue systems, the chitchat has no predefined goals and often targets human-human dialogues where the underlying goal is often not known or hard to define in advance such as reserving a table at a restaurant or booking a flight. Recent advancements to neural response generation address this problem by grounding systems in textual knowledge sources such as Foursquare (Ghazvininejad et al. 2017), the users or agents visual environment (Das et al. 2017; Mostafazadeh et al. 2017), and persona or emotion of the users (Li et al. 2016; Al-Rfou et al. 2016; Huber et al. 2018). All these works essentially try to augment their conversational models to not only represent the dialogue history, but also contextual information drawn from the dialogue scenes, such as an image (Das et al. 2017; Mostafazadeh et al. 2017) or textual information (Ghazvininejad et al. 2017).

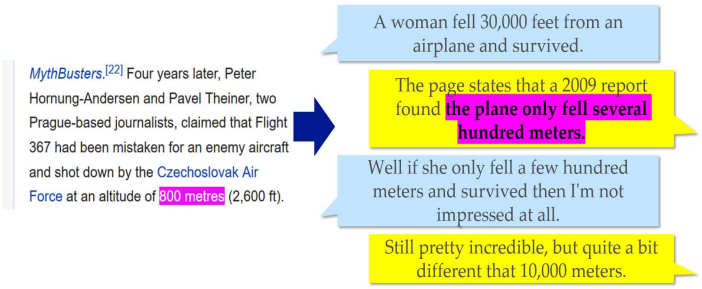

Figure 1: A sample grounded conversation at Reddit (Gao, Galley, and Li 2018a)

The 7th Dialog System Technology Challenge (DSTC7) track 2 proposes an E2E conversational modeling task, where the goal is to generate conversational responses that go beyond chitchat, by injecting informational responses that are grounded in external knowledge bases such as Foursquare, Wikipedia, Goodreads, or TripAdvisor (Yoshino et al. 2018). Figure 1 presents a sample grounded

*Corresponding author: chen.xin@acm.org

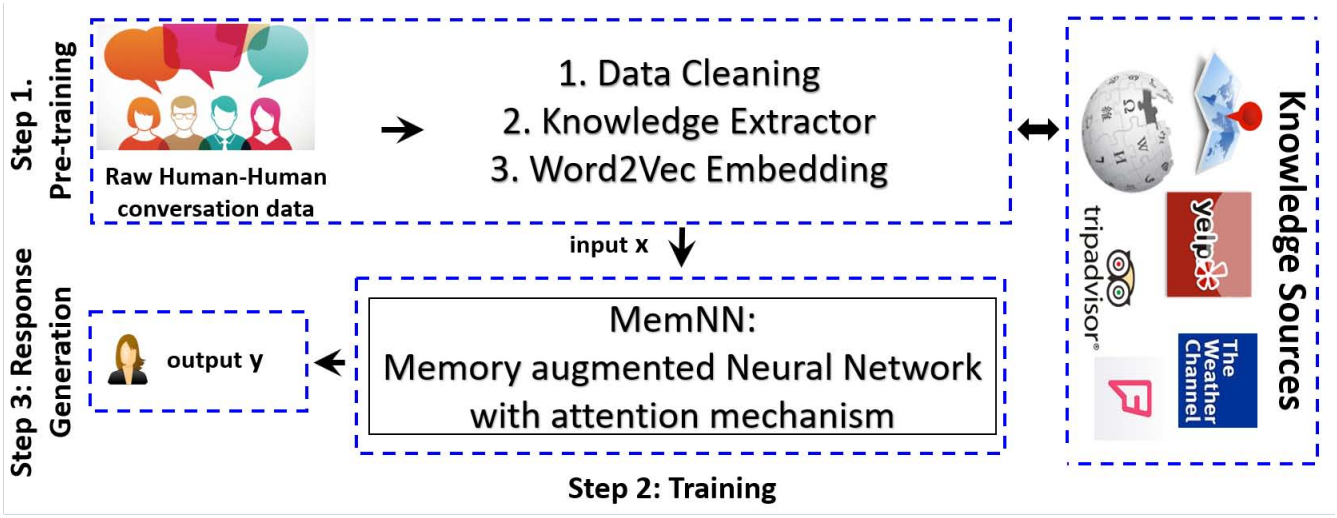

Figure 2: The OneConn-MemNN System

conversation at Reddit (Gao, Galley, and Li 2018b). In this paper, we present our OneConn-MemNN system participated in DSTC7 Track 2, which aims at generating knowledge-grounded conversational responses in a fully data-driven manner.

To summarize, our proposed system has the following key features and contributions:

- It tailors a memory augmented network with attention mechanism to not only represent the human-to-human conversations but also external knowledge sources. We integrate this model to achieve the highest score on "Appropriateness" [1] and the second highest score on "Informativeness & Utility" [2] by human evaluation.
- It targets open-domain E2E response generation using the large-scale Reddit dataset, which contains more diverse topics, user backgrounds, and more natural conversations as opposed to other chitchat datasets such as Twitter or Ubuntu.
- It enriches the generated responses using entities and factual contents drawn from contextually-relevant Wikipedia snippets of text, without explicit slot filling.

The remainder of the paper is organized as follows: **Section 2** introduces the framework of the OneConn-MemNN system. In **Section 3**, we introduce the memory augmented network with attention mechanism. In **Section 4**, we conduct experiments in the Reddit dataset to show the effectiveness of our system. We summarize our work in **Section 5**.

2 The OneConn-MemNN System

Figure 2 shows a high-level overview of the OneConn-MemNN system. The input to the system consists of conversational data and grounded knowledge facts. The whole system comprises 3 components: the preprocessing part, the training part, and the response generation part. A pre-training component first preprocesses the conversation data and extracts facts that are relevant to context of the conversation. Before the training phrase, pretrained word embeddings are used to initialize the first layer of the model for the following response generation. The pre-processed conversations and facts are fed into our customized E2E neural network based on sequence-to-sequence (SEQ2SEQ) models. Finally, the trained model could automatically generate responses either through offline batch processing or in an online mode that can interact with real human beings.

2.1 Preprocessing

Before training, preprocessing is the key to achieve a good performance for most of the systems. We believe the preprocessing is crucial for our system to generate reasonable human-like responses. The particular thresholds or hyperparameters are chosen based on the exploratory data analysis of the datasets used in DSTC7 Track 2, but the idea could be easily generalized to other corpura.

Data Cleaning It is important to purify the conversational sentences in terms of semantic structure, especially in the case that the training data is collected from the Internet sources such as Reddit or Twitter. There are a large number of URL links, which cannot be treated as words because of the extremely low frequency. On the other hand, the hyperlinks could be semantically meaningful, so we cannot simply remove them either. Our approach is to replace all the URL links with a unified token "URL" to represent them in a generalized level. Another source of noises comes from special syntax, such as Markdown, hashtags, emoticons, etc. A simple way to eliminate these noises is to remove all special symbols and only keep alphabetic characters, numbers and basic punctuation marks. Our experiments during system development reveal that repeating tokens in training data

[1] This evaluation criterion asks the crowdsourced judges whether the system response is conversationally appropriate and relevant.

[2] This evaluation criterion asks the crowdsourced judges whether the system response is interesting and informative.

could be detrimental to a SEQ2SEQ model. For example, the model might frequently predict another period after a period, which results in a string of periods. To resolve this, we remove repeating punctuation characters in the conversational data. To accelerate the training convergence, the lengths of sentences should be limited in a certain range. Instead of using the entire conversation history, we keep the most recent k turns as the context. We further trim the extremely long sentences to ensure our training sentences are within a moderate length range.

Most of the ideas to clean conversational data also apply to grounded facts, because the contextually-relevant facts are typically unstructured texts from external knowledge sources in the Internet, such as Wikipedia or Foursquare. In particular, in DSTC7 Track 2 task, the grouned facts are snippets consisting of almost all the texts from a given source page, within which only certain a few sentences might be informative. We assume that informative facts are located in those sentences with adequate length, whereas snippets containing fewer words are mostly "garbage" sentences. Following this assumption, we keep only fact snippets with word counts greater than a specified threshold, and trim all cut short ones. A further fine filtering is achieved by the Knowledge Extractor which will be discussed below.

Knowledge Extractor The set of facts that are relevant to context of the conversation is provided and fixed, because to retrieve the contextually-relevant facts from a large pool is not the focus of this sentence generation task. However, the specific set of relevant facts could change during the progress of a conversation. To use the most relevant facts for each context-response pair, top-K (here we choose K = 10) factual snippets are retrieved based on TF-IDF similarity between context and fact.

Word2Vec Embedding The data gathered from Reddit, similar to other internet sources such as Twitter, contains a large number of abbreviations, slang and typos, potentially yielding more noises. To eliminate the effects of the diverse phraseology, we use training data to pre-train a Word2Vec model, which is afterwards used to initialize the embedding weights of our system.

2.2 Training

During the training phase shown in Figure 2, we construct a memory augmented SEQ2SEQ model with attention mechanism and train the model using the Cross-Entropy (CE) criterion and the ADAM optimizer (Kingma and Ba 2014).

Sequence-to-Sequence Models Our conversation modeling uses an adapted SEQ2SEQ models (Cho et al. 2014b; Shang, Lu, and Li 2015; Sordoni et al. 2015; Vinyals and Le 2015), as shown in Figure 3. The general framework of a SEQ2SEQ model maps one sequence to another using a neural network. It uses an encoder to take in the sequence of words in conversational history and outputs a fix-sized context vector which summarizes previous dialogue histories. It then uses a decoder to generate the response word-by-word based on the recurrent hidden state and the context vector. The encoder and decoder process the input sequence and output sequence based on recurrent neural network (RNN). The encoder RNN unit takes in a word and updates its hidden state while decoder RNN is computed by a nonlinear activation function named Gated Recurrent Unit (GRU) (Cho et al. 2014b). The output distribution of the decoder is used to predict the next token and is parameterized by a softmax function over an affine transformation of the decoder RNN hidden state. The model parameters are turned by maximizing the loss function over the training instances by stochastic gradient descent. In the section 3, we explain in detail our customized SEQ2SEQ model.

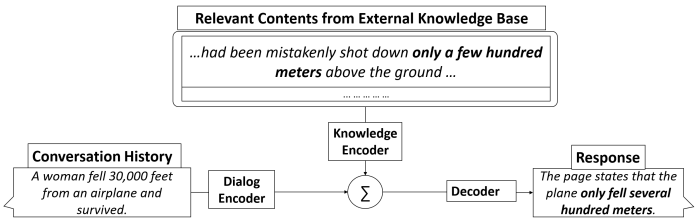

Figure 3: Memory Augmented Network with Attention Mechanism

The Objective Function In this paper, the adapted SEQ2SEQ model is trained with the Cross-Entropy (CE) criterion using the training corpus. We define the loss function as the average negative log likelihood, with teacher forcing and the gradient clipping heuristic to aid the convergence (Goodfellow, Bengio, and Courville 2016).

Optimization Before training, we initialize the individual encoder with Word2Vec embeddings and then optimize the loss function with gradient descent methods. Given the data and models, the network parameters get updated with the ADAM optimizer, with the checkpoint enabled to save the model parameters for either response generation or continuing training from where left off.

2.3 Response Generation

After training a model, we want to be able to talk to the bot ourselves, the following generation phase employs model-based sentence generation.

Greedy Decoding First, we must define how we want the model to decode the encoded input. Greedy decoding is the decoding method that we use for generating responses from trained models. For each time step, we simply choose the word from decoder output with the highest softmax value. This decoding method is optimal on a single time-step level.

Interactive Mode and Batch Processing With the decoding method defined, the user can interact with the trained models embedded in the system. When called, an input text field will spawn in which we can enter our query sentence. After typing the input sentence and pressing Enter, our text is normalized in the same way as our training data, and is ultimately fed to generate a decoded output sentence. For evaluation purpose, we can also run offline batch processing to generate lots of responses simultaneously.

3 Memory Augmented Network with Attention Mechanism

The brain of our neural conversation model is based on a SEQ2SEQ mapping process using Bidirectional recurrent neural networks (RNNs). The goal of a SEQ2SEQ model is to take a variable-length sequence as an input, and return a variable-length sequence as an output using a neural network model. In particular, we develop a memory augmented network with attention mechanism that is able to capture both the human-to-human conversations and entities and factual contents drawn from external knowledge sources, as shown in Figure 3. Note that we take the Reddit dataset and the Wikipedia as a case study for knowledge-grounded conversation modeling, but the framework can be generalized to a wide range of datasets.

In the following, we will describe **Encoder RNN**, **Memory Encoder**, and **Decoder RNN with Attention**.

3.1 Encoder RNN

The basic component of our encoder is a bidirectional variant of the multi-layered GRU, invented by (Cho et al. 2014b), which gives us the advantage of encoding both past and future context.

Essentially, the bidirectional GRU has two independent RNNs. One RNN is fed the input sequence in normal sequential order, and the other RNN is fed the input sequence in reverse order. The outputs of each RNN are summed at each time step. There is another hidden embedding layer to encode the word sequence in an arbitrarily sized feature space. When trained with large-scale real-world corpus, these hidden embedding layers should encode semantic similarity between similar meaning words.

Given an input sequence of tokens, the encoder RNN iterates one token at a time and generates an output vector and a hidden state vector. In the next iteration, the output vector is recorded and the hidden state vector is reused. The encoder and decoder RNNs map the context of each word in the sequence into a set of points in a high-dimensional space, which is then the learned representation learned for each word. For a padded batch of sequences, the model need to pack and unpack paddings around the RNNs and pass padded sequence and packed sequence respectively.

3.2 Memory Encoder

Memory network models are widely used in natural language processing tasks to make grounded based inferences (Weston, Chopra, and Bordes 2014). The Knowledge Encoder in figure 3 is a simplified version of the Memory Network model proposed by (Sukhbaatar et al. 2015; Weston, Chopra, and Bordes 2014). To model entities and factual contents mentioned in the conversation, our memory augmented model simply adds one extra layer of memory encoder. This associative memory encoder maps the external facts into vectors and then adds up the vectors transformed from the dialog encoder. Finally, the model retrieves and weights these facts based on the conversation history to generate a knowledge-grounded response.

3.3 Decoder RNN with Attention

The decoder RNN generates the response sentence one token at a time, based on the encoders context vectors and internal hidden states. It continues generating the next word until it outputs a stopping symbol that represents the end of the sentence.

However, a basic SEQ2SEQ decoder does not account for the information loss, especially when dealing with long input sequence, and then significantly limits the real-world applications of such model. To alleviate this information loss issue, we introduce an attention mechanism that allows the decoder to overlook the whole context and only pay attention to selected parts of the context sequence (Bahdanau, Cho, and Bengio 2014; Luong, Pham, and Manning 2015). Such attention mechanism relies on the decoder's hidden state and the encoder's outputs. The decoder generates attention weights and gets multiplied by the encoder outputs, giving us a weighted sum which indicates a refined part of the encoder the model should pay attention to.

4 Experiments

To evaluate the performance of our system, we conduct experiments on Reddit datasets. The Reddit datasets, consisting of conversational data and grounded facts, are generated by official scripts from DSTC7 track 2. Table 1 shows the basic statistics of the datasets.

	Train data	Dev. set	Test set
# dialogue turns	2,364,239	119,478	13,440
# facts	15,181,673	1,675,056	582,944
# tagged facts	2,288,351	369,423	139,406

Table 1: Data statistics of Reddit and knowledge sources

4.1 Conversational Data from Reddit

Reddit is a social media source that is also practically unbounded, and represents about 3.2 billion dialogue turns as of July 2018. It was for example used in (Al-Rfou et al. 2016) to build a large response retrieval system. Reddit data is organized by topics (i.e.subreddits), and its responses dont have a character limit as opposed to Twitter.

We preprocess the conversational data from Reddit before the model training phrase. To eliminate the noises, we simply remove all non-letter characters except the basic punctuation marks. We also replace all the URL links with a unified token "URL", and remove repeating punctuation characters. Repeating tokens in training data could be detrimental to a SEQ2SEQ model, because the model might go to a infinite loop during decoding.

To simplify the learning tasks of our model, we keep the most recent 1 turn rather than using the entire dialog history as input. The problem of this approach is the model would omit some important information within the context, which, however, will be compensated during fact grounding phrase. We will discuss this more in the results.

Finally, we trim our context-response pairs in two steps to further eliminate some noises. 1. Long sentences with more

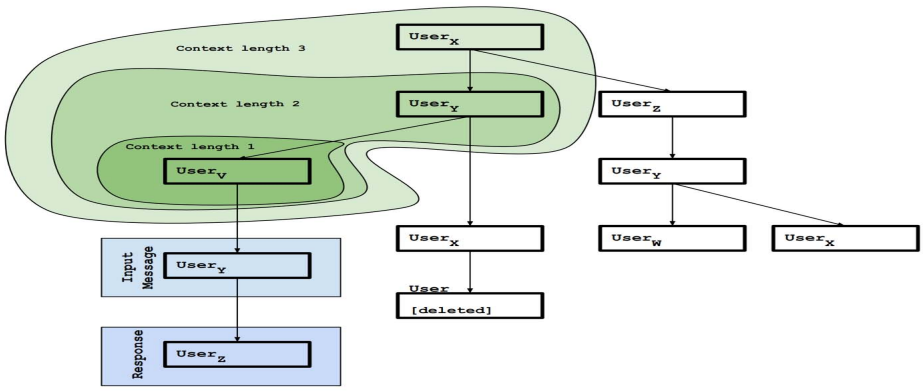

Figure 4: A diagram of the Reddit comment tree structure (Reddit Post). UserZ responded to the message produced by UserY (blue). If we follow the ancestors of the input message, we can construct several contexts of different lengths (green). (Al-Rfou et al. 2016)

than 60 words are trimmed, which consists of 15.76% of the original sample size. The cut-point 60 is chosen based on the observation that after keeping the most recent 1 turn in context, most of the sentences containing words less than 60. In the plotting of word count distribution, there is a obvious elbow point around 60. 2. Sentences containing rare words are trimmed. We define rare words as those words with a frequency lower than 7 in our training corpus. The benefit of trimming rare words is considerable: the vocabulary size is cut down from 255,711 to 59,233 (77% off), which would significantly save the time and space costs during training process. After the two-step trimming, there are 1,617,832 training samples left (68.4% of original sample size). We still have enough data for training.

4.2 Grounded Facts from Knowledge Sources

Following the assumption that informative facts are located in those sentences with adequate length, we trim the fact dataset by keeping snippets with no less than 5 words after removing all the HTML tags and punctuation characters. After filtering, 2,288,351 snippets remain in the training fact set (15.1% of the original number 15,181,645, which is slightly different from the official statistics shown in Table 1). The threshold 5 is chosen arbitrarily, because the purpose of this preliminary filtering is to eliminate most potential garbage snippets. Further selecting relevant facts is achieve by choosing the top 10 facts with the highest TF-IDF scores (Ghazvininejad et al. 2017).

Method	BLEU	NIST	METEOR	Entropy	Diversity
Baseline (constant)	2.87%	0.184	7.48%	1.609	0.000
Baseline (random)	0.86%	1.637	5.91%	10.467	0.647
Baseline (seq2seq)	1.82%	0.916	6.96%	5.962	0.048
Human	3.13%	2.650	8.31%	10.445	0.670
Attention	1.32%	2.124	6.81%	7.206	0.124
OneConn-MemNN	**1.32%**	**1.515**	**6.43%**	**7.639**	**0.171**

Table 2: Model Comparison

4.3 Model Comparison

The outputs are evaluated based on five automatic evaluation metrics as well as two human evaluation scores. In particular, **BLEU**, **NIST**, **METEOR**, **Entropy**, and **Diversity** are used for automatic evaluation. For human evaluation, a per-response judging is performed via Mechanical Turk based on two criteria (i.e., relevance and informativeness) on a 5-point Likert scale. Our system is compared with several official baselines as well as human responses. We also include a fact-ungrounded SEQ2SEQ model with attention mechanism (Attention) into model comparison, as an external aid to explain the effectiveness of our OneConn-MemNN system.

Results We first report the model performance regarding the automatic evaluation metrics of BLEU, NIST, METEOR, Entropy and Diversity for the OneConn-MemNN against other baselines. BLEU, NIST, METEOR are metrics that are used to evaluate relevance on response generation task and have shown reasonable correlations with human judgments. Although Human response achieves the highest performance in all three metrics, the results for Baseline (constant) and Baseline (random) are not consistent. For example, the BLEU of constant baseline is significantly higher than that of random baseline, while the NIST is much lower in constant than in random. Furthermore, our system and seq2seq baseline are always somewhere between Baseline (constant) and Baseline (random). The results indicate the potential limitations of current automatic relevance metrics. We calculate the correlations between these three metrics and human relevance evaluation based on the baselines, human response and primary submissions from six team in DSTC7 Track 2. The correlations with human judged relevance score are 0.59, 0.61 and 0.71 for BLEU, NIST, and METEOR, respectively. The moderate correlations imply none of these metrics along could be enough used for cross validation tasks such as model selection or hyper-parameter tuning. A more dedicate metric could potentially accelerate the progress in end-2-end response generation system.

Entropy and Diversity metrics measure the overall di-

Context	Responses
Arsenal becomes the most winning team in the fa cup with 13 wins.	**SEQ2SEQ Attention:** I m not going to say that but i m not going to say that. **OneConn-MemNN:** I m not sure arsenal is the most dominant team in the league.
During wwi , king albert i of belgium personally led the defense against the germans , while the queen served as a nurse , and their son enlisted in the infanty.	**SEQ2SEQ Attention:** I don t think i ve ever heard of this. **OneConn-MemNN:** I m sure he was a monarchist.
That the empire state building only took a little over a year for its construction to complete (410 days twelve days ahead of schedule).	**SEQ2SEQ Attention:** I m not sure why i m not a big deal . i m not sure why it s a year old. **OneConn-MemNN:** s I m a big fan of the empire building.
A hiroshima policeman went to nagasaki to teach other police officers to duck and cover in the days between the bombings . not a single officer died in the nagasaki blast.	**SEQ2SEQ Attention:** I m not sure if this is a joke. **OneConn-MemNN:** The policeman died in hiroshima.

Table 3: Response examples by the OneConn-MemNN system and the baseline

versity of the responses, where constant baseline achieves a very low score. Our system gets a moderate diversity. The Memory Network diversifies the response (note the increases from Attention to OneConn-MemNN).

Our system ranks top 1 for relevance and second for informativeness. Case study shows that our knowledge-grounded responses significantly increase the relevance. Instead of chitchats generated by Attention model, our system outputs contain highly relevant contents. However, these contents are difficult to find according reference in fact dataset. More likely, these relevant responses are learned from training data, the Memory Network functions as another way of attention to help our model to pick them out. This finding is interesting, because originally we introduce Memory Network to our system in order to enable it to leverage external knowledge for generating responses. It's aimed to increase the informativeness rather than relevance. One possible way to interpret this interesting phenomena is the Memory Network uses bag-of-words representations of facts, which potentially losses some semantic information in the sentence level. Because the top K related facts are selected based on tf-idf similarities compared to context, relevant words will be emphasized. As a result, using the output of Memory Network to update the hidden state of Decoder would work as an different kind of attention mechanism to guide Decoder to choose more relevant response.

4.4 Case Study

Table 3 presents examples of generated responses for our OneConn-MemNN system and the baseline model with attention mechanism. It is clear that the OneConn-MemNN system performs better at understanding contexts. The responses of the OneConn-MemNN system are not only more appropriate but also more informative and useful. The baseline tends to make generic, safe but dull responses, while the OneConn-MemNN system tends to output more specific answers with entities or factual contents extracted from the context such as "the most dominant team", "monarchist", "the empire building", and "hiroshima".

5 Conclusion

In this paper, we have studied the E2E approaches for conversation modeling and have focused on how to generate the responses that are both relevant and the informative. To achieve the goal, we have proposed the OneConn-MemNN system that utilizes a memory augmented SEQ2SEQ model to generate knowledge-grounded conversational responses. We achieve this by adding an extra memory encoder layer to incorporate the contextual entities or factual contents from external knowledge base. We also enable the attention mechanism to represent the human-to-human dialogue within and across the conversation histories. We experiment the OneConn-MemNN system on a large-scale conversation dataset extracted from Reddit and associated Wikipedia links in a domain-agnostic manner. Both the automatic evaluation metrics and human evaluation scores have verified the effectiveness of our proposed system.

References

Al-Rfou, R.; Pickett, M.; Snaider, J.; Sung, Y.-h.; Strope, B.; and Kurzweil, R. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014b. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Gao, J.; Galley, M.; and Li, L. 2018a. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1371–1374. ACM.

Gao, J.; Galley, M.; and Li, L. 2018b. Neural approaches to conversational ai. In *Advances in neural information processing systems*. ACL and SIGIR tutorial.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gu, J.; Lu, Z.; Li, H.; and Li, V. O. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Huber, B.; McDuff, D.; Brockett, C.; Galley, M.; and Dolan, B. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 277. ACM.

Kalchbrenner, N., and Blunsom, P. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 48–54. Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mei, H.; Bansal, M.; and Walter, M. R. 2017. Coherent dialogue with attention-based language models. In *AAAI*, 3252–3258.

Mostafazadeh, N.; Brockett, C.; Dolan, B.; Galley, M.; Gao, J.; Spithourakis, G. P.; and Vanderwende, L. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Och, F. J., and Ney, H. 2004. The alignment template approach to statistical machine translation. *Computational linguistics* 30(4):417–449.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, 3776–3784.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Shao, L.; Gouws, S.; Britz, D.; Goldie, A.; Strope, B.; and Kurzweil, R. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. *arXiv preprint arXiv:1701.03185*.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, 2440–2448.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, 2692–2700.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks.

Yao, K.; Zweig, G.; and Peng, B. 2015. Attention with intention for a neural network conversation model. *arXiv preprint arXiv:1510.08565*.

Yoshino, K.; Hori, C.; Perez, J.; D'Haro, L. F.; Polymenakos, L.; Gunasekara, C.; Lasecki, W. S.; Kummerfeld, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B.; Gao, S.; Marks, T. K.; Parikh, D.; and Batra, D. 2018. The 7th dialog system technology challenge. *arXiv preprint*.