MODELING CONVERSATION CONTEXT BY ADAPTING COGNITIVE ARCHITECTURES

Sashank Santhanam & Samira Shaikh
Department of Computer Science
University of North Carolina at Charlotte
{ssantha1,samirashaikh}@uncc.edu

ABSTRACT

In this paper, we present an approach for language generation for open-domain dialog systems inspired by neurocognitive memory processes. We overcome the drawbacks of the traditional *seq2seq* architecture by augmenting the architecture with two types of memory, namely, long-term and working memory based on the Standard Model of cognition. We also implement a novel action selection mechanism that helps identify the relevant utterances containing salient information from long-term memory to working memory. To evaluate our model, we compare our action selection mechanism with the state-of-the-art baseline and observe improvements in the identification of most salient utterances over long conversations and our mechanism shows a higher correlation to human rankings.

1 INTRODUCTION

Extant approaches to natural language generation have typically been formulated as sequence-to-sequence (*seq2seq*) frameworks, an adaptation of machine translation systems (Vinyals & Le, 2015; Sutskever et al., 2014). Prior research has shown that engaging with these systems for longer interactions could result in dull and generic responses due to their reliance on the last utterance in the dialogue history as contextual information (Tian et al., 2017). To make better use of context, researchers have used both hierarchical and non-hierarchical models (Tian et al., 2017), but these models still suffer from sub-optimal performance due to the inclusion of entire dialogue history that may contain irrelevant utterances (Wang et al., 2018). Thus, we formulate the problem as: *How do we encode longer context into the NLG system such that the algorithm can focus on the salient information in the conversation (e.g. important topics, entities) while appropriately discounting parts of conversation that may primarily serve to preserve social conventions (e.g. ”ah”, ”ok” etc.)?*

To address this challenge, we adapt the Standard Model (Norris, 2017), an established model of memory in cognitive science. This model provides the framework with which to conceptually and practically address both long-term memory and short-term memory (working memory) - to incorporate the longer context of conversation along with the immediate context. The model also provides an action-selection mechanism acting as a bridge between the long and short-term memory. To the best of our knowledge, our work is the first to use the Standard Model of Cognition to more closely tie the NLG system to the way human cognition works. Our work makes the following contributions:

1. A novel **cognitively-inspired natural language generation model** that accounts for larger contexts while appropriately discounting irrelevant utterances from the context through *long-term* and *working memory*.
2. Incorporating **action selection mechanism**, an adaptation of attention mechanism, to help identify salient context from Long-Term Memory.

We find that our approach is better able to identify contextual utterances that show high correlation with human rankings and outperform state-of-the-art performance on longer conversations.

2 RELATED WORK

Cognitive architectures identify the structures and processes in the brain and facilitate understanding the interactions between them (Newell, 1994; Sun, 2007). More concretely, cognitive architectures help to understand how perception, vision, action selection along with the ability to store knowledge using memories (short and long-term) make agents function with human-level intelligence (Langley et al., 2009; Kotseruba & Tsotsos, 2016). ***Memory*** plays a significant role in cognitive architectures. In the Standard Model architecture, Declarative Memory stores episodic knowledge as part of Long-term memory which is relevant to our generation task. The short-term memory (working memory) works along with an action selection mechanism to retrieve, store and process relevant pieces of information for a particular task. Attention mechanism helps retrieval of relevant information, a form of ***action selection*** within cognitive architectures (Bahdanau et al., 2014). The concept of memory, conceptualized as encoding contextual information in dialogue, has previously been explored in question-answering systems (Sukhbaatar et al., 2015; Kumar et al., 2016). However, the usage of memory networks for dialogue generation is still in its infancy.

The area of ***dialogue systems using deep learning*** has been studied extensively, both in the open domain (Niu & Bansal, 2018; Rashkin et al., 2019) and goal-oriented situations (Lipton et al., 2018). Sequential and hierarchical models (Serban et al., 2016) have been proposed to make better use of context. Our work is most closely related to Tian *et al.* (2017), who demonstrated that hierarchical methods perform better than non-hierarchical models in encoding context. However, they focused on a single prior utterance as context and did not account for non-informative utterances. In contrast, we propose an action-selection to identify multiple informative, salient context utterances.

3 Model Architecture and Corpus

Our model is shown in Figure 1 - called the Cognitive Memory Architecture (CMA) - a dual memory augmented encoder-decoder model inspired by the Standard Model. The model comprises of the following components:

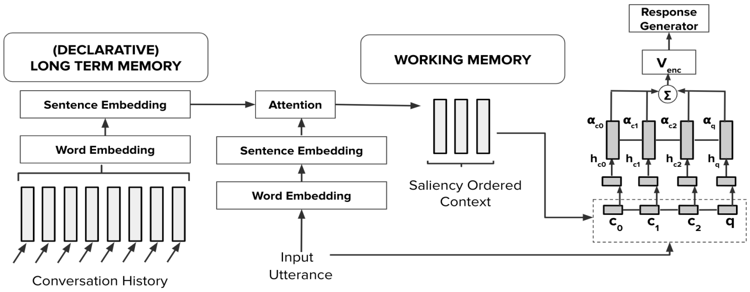

Figure 1: Architecture of CMA model with two memory components namely, long-term and working memory adapted from the Standard Model of Cognition that augments the input utterance in an hierarchical fashion.

Long Term Memory, which stores the history of the conversation. The input to the long-term memory is the historical sequence of utterances U_t where *1≤t≤8*. We make a simplifying assumption that only a maximum of 8 utterances is stored in long-term memory (we address this in Section 5). In our approach, we convert each utterance in the history (U) and input utterance (Q) to its respective sentence embedding as the sum of the word embeddings (Eq. 1). Next, we compute the cosine similarity between the sentence embeddings of each of the utterances in the dialogue history e_{u_i} and the sentence embedding of input utterance e_q, where e_w represents the embedding of the word.

$$e_{u_i} = \sum_{w \in u_i} e_w, \quad \& \quad e_q = \sum_{w \in q} e_w \tag{1}$$

We improve similarity calculation by introducing two additional parameters: λ represents the order of the sequence (in the range of $1 \leq \lambda \leq n$ in our work and n represents the length of the dialogue history) and τ is counter balancing parameters and is the ratio between the number of words in the utterance to the maximum length of the target utterance. Our intuition in defining these additional parameters in Eq. 2 is that (a) the **most important utterances** to focus on when generating a response may be present in **early part of dialogue history** and (b) these utterances should be given

relative importance that is meaningfully encoded when generating the response.

$$s_{u_i} = sim(u_i, q) + \lambda \cdot \tau, \quad \& \quad sim(u_i, q) = \frac{e_{u_i} \cdot e_q}{||e_{u_i}|| \cdot ||e_q||} \tag{2}$$

Action Selection - To obtain the relative importance scores of each utterance in the history to the query utterance, we perform action selection using Bahdanau-style attention mechanism (2014):

$$\alpha_{u_i} = \frac{exp(s_{u_i})}{\sum_{i=0}^{t} exp(s_{u_i}) + exp(s_q)}, \quad \& \quad \alpha_q = \frac{exp(s_q)}{\sum_{i=0}^{t} exp(s_{u_i}) + exp(s_q)} \tag{3}$$

where α_{u_i} and α_q represents the importance score and s_q will be 1 as the query utterance similarity is computed against the same vector and t is the number of utterances in the dialogue history.

Working Memory - stores the utterances necessary for generating an appropriate response while appropriately down-weighting the non-useful utterances. The working memory stores C_n utterances that have the highest score produced by the Action Selection Mechanism (Equation 3). The ***Input Module*** and ***Response Module*** - follow the Hierarchical Encoder-Decoder architecture (Serban et al., 2016). We calculate the final vector as a sum of the hidden vectors weighted by their saliency value α_u of the utterances in the working memory and α_q obtained from above Equation 3.

Our model is developed and tested on the MovieTriples corpus from Serban *et al.* (2016). The MovieTriples corpus contains triplets of dialogues. We merged all triplets for a given movie and then divided the resulting corpus into conversation sequences of 10 utterances. Each sequence is further divided into dialogue history (of length $\leq$ 8), query and target utterances. After pre-processing, the dataset comprised of 42738 conversations in training and 1000 conversations in test set. The average length of dialogue history in the training set was 5.5 and 5.24 in the test set.

4 Experiments and Results

We hypothesize that salient information is present in the earlier context of the conversation and cannot be obtained from only using one or two most recent utterances as context. To obtain quantitative evidence to test our hypothesis, we recruited 60 annotators to annotate 120 randomly sampled conversations from our test data.[1] We asked each annotator to rank order each utterance from the dialogue history in order of its saliency. We used the intra-class correlation coefficient scores to measure the inter-rater reliability for more than two raters (Shrout & Fleiss, 1979). We find that the rankings obtained from the user study have a **high consistency and agreement value of 0.88** (both values are statically significant with p-value<0.001).

To understand the impact of longer conversational context and the saliency of historical utterances in the task of response generation, we grouped the conversations by length of the dialogue history into 3 groups: (1) **Long** - Dialogue histories of length $\geq$ 6; (2) **Medium** - Dialogue histories of length 4 and 5; (3) **Short** - Dialogue histories of length $\leq$ 3. We also notice that from the randomly sampled 120 conversations, **91** conversations had a Long dialogue history, **16** had a Medium dialogue history and **13** had a short dialogue history. Table 1 shows the importance of recent utterances when compared to the historical utterances in the conversation history as rated by the human annotators. *Recent Utterances* refer to the two utterances immediately prior to the query utterance. *Earlier Utterances* refer to the utterances in conversation history beyond the two utterances prior to query utterance. **Finding 1:** We observe the following from the human annotators' rankings: 1) For long and medium conversation histories, only around 63% , the recent utterances are labelled as the most salient in capturing the context of the conversation. This means that around 36% of the time, salient information is included in history of dialogue beyond the immediate one or two utterances prior to the query utterances. 2) For short conversations, we find that 91.67% of the time the historical utterances in conversation history are referred to as most important.

Next, we compare our results with an existing state-of-the-art model proposed by Tian *et al.* (2017) as their work is the most similar work to ours, in that it also attempts to identify and encode context importance. Figure 2 shows the accuracy of our CMA model and Tian *et al.* model in identifying the top three salient contexts (as judged by human annotators). **Finding 2:** We demonstrate using Figure

[1] Instructions and sample questions given to annotators are presented in supplemental materials

Type of Dialogue History	Recent Utterances	Earlier Utterances
Long (n=91)	63.95%	36.05%
Medium (n=16)	64.29%	35.71%
Short (n=13)	8.33%	91.67%

Table 1: Importance of recent utterances when compared to earlier utterances present in dialogue history.

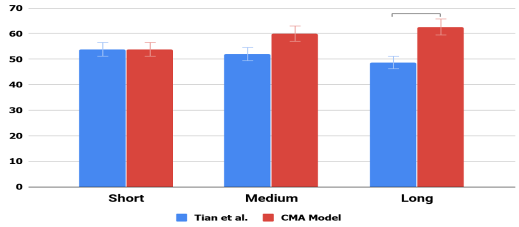

Figure 2: Accuracy of identifying salient utterances by action selection mechanism in our proposed CMA model compared to state of the art (Tian et al., 2017).

2 the ability of the CMA model to identify salient utterances and outperform the existing state-of-the-art methods in longer dialogue histories (statistically significant *p<0.001* while achieving comparable performance in shorter dialogue histories.

Additionally, Table 2 shows Spearman and Kendall Tau correlation results between the action-selection mechanism and human rankings (HR). **Finding 3:** We observe that CMA action-selection mechanism shows significantly higher correlation to human judgments than the method proposed by Tian *et al* (2017) over Long and Medium dialogue histories. This is likely because of the addition of parameters λ and τ that provide additional importance based on the length of sequence and order of the sequence. However, we do notice that method proposed by Tian *et al.*(2017) outperforms our method on short dialogue histories (although this difference was not statistically significant).

		Long	Medium	Short
Spearman	CMA∼HR	**0.46*****	**0.27***	0.13
	Tian *et al.*∼HR	0.33***	0.12	**0.25**
Kendall Tau	CMA∼HR	**0.37*****	**0.23***	0.12
	Tian *et al.*∼HR	0.26***	0.10	**0.24**

Table 2: Spearman and Kendall Tau Correlation between CMA Action Selection mechanism and Tian *et al.* method to the rankings from human annotators. *** $p < 0.001$; * $p < 0.05$; (HR: Human Ranking)

Model	BLEU Score	Diversity	Length
CMA	**0.091*****	0.00089	**9.24*****
Context Seq2Seq	0.060	**0.00091**	6.76
No Context Seq2Seq	0.063	**0.00091**	6.56

Table 3: Performance of CMA model and baselines on BLEU score, diversity, length and Gunning-Fog Index.
*** $p < 0.001$

We report performance of the CMA model on dialogue generation using traditional metrics such as BLEU, Diversity, and Length consistent with existing literature in Table 3. In Table 3, **CMA** refers to the model that uses salient contexts identified though action selection mechanism, **Context Seq2Seq** model uses the most recent utterance as the context, and **No Context Seq2Seq** refers to model that uses no context. We use the Distinct-1 metric to calculate diversity as the number of distinct unigrams over the total number of generated tokens (Li et al., 2016). **Finding 4:** The CMA model is less diverse than the baselines (although no statistical significance); but able to generate longer, coherent sentences and significantly outperform baselines on BLEU score and Length.

5 Conclusion and Discussion

We have shown how the long-term and working memory as described by cognitive architectures can be adapted to augment *seq2seq* models for dialogue generation.We find that the action selection mechanism is able to identify salient utterances and outperform extant methods to maintain the conversation context. We make a simplifying assumption that long-term (declarative) memory has at most 8 prior utterances for practical reasons such as training time and compute resources. In ongoing improvements, we are adapting our models to incorporate more context, include world knowledge (e.g. Wikipedia), and use transformer architectures (e.g. BERT) (Devlin et al., 2019).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September 2014. URL `https://arxiv.org/abs/1409.0473`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Iuliia Kotseruba and John K Tsotsos. A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications. *arXiv preprint arXiv:1610.08602*, 2016.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pp. 1378–1387, 2016.

Pat Langley, John E Laird, and Seth Rogers. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160, 2009.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119. Association for Computational Linguistics, 2016. doi: 10.18653/v1/N16-1014. URL `http://www.aclweb.org/anthology/N16-1014`.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Allen Newell. *Unified theories of cognition*. Harvard University Press, 1994.

Tong Niu and Mohit Bansal. Polite dialogue generation without parallel data. *Transactions of the Association of Computational Linguistics*, 6:373–389, 2018.

Dennis Norris. Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9):992, 2017.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pp. 3776–3784, 2016.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pp. 2440–2448, 2015.

Ron Sun. The importance of cognitive architectures: An analysis based on clarion. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(2):159–193, 2007.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pp. 231–236, 2017.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. Chat more: Deepening and widening the chatting topic via a deep model. 2018.