Memory-Based Deep Neural Attention (mDNA) for Cognitive Multi-Turn Response Retrieval in Task-Oriented Chatbots

Article *in* Applied Sciences · August 2020

DOI: 10.3390/app10175819

3 authors, including:

Obinna Agbodike
NAU
6 PUBLICATIONS **10** CITATIONS

SEE PROFILE

Lei Wang
AKKA Technologies
175 PUBLICATIONS **2,860** CITATIONS

SEE PROFILE

applied sciences

MDPI

Article

Memory-Based Deep Neural Attention (mDNA) for Cognitive Multi-Turn Response Retrieval in Task-Oriented Chatbots

Jenhui Chen [1,2,3,4,†], Obinna Agbodike [5] and Lei Wang [6,*]

1 Department of Computer Science and Information Engineering, Chang Gung University, Kweishan, Taoyuan 33302, Taiwan; jhchen@mail.cgu.edu.tw
2 Artificial Intelligence Research Center, Chang Gung University, Kweishan, Taoyuan 33302, Taiwan
3 Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Kweishan, Taoyuan 33375, Taiwan
4 Department of Electronic Engineering, Ming Chi University of Technology, Taishan District, New Taipei City 24301, Taiwan
5 Department of Electrical Engineering, Chang Gung University, Kweishan, Taoyuan 33302, Taiwan; d0721009@cgu.edu.tw
6 School of Software, Dalian University of Technology, Dalian 116024, China
* Correspondence: lei.wang@dlut.edu.cn
† This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-182-042-MY2; and in part by the Chang Gung Memorial Hospital, Kweishan, Taoyuan, Taiwan, under Grants CMRPD2J0012 and CMRPD2I0052.

Received: 28 July 2020; Accepted: 21 August 2020; Published: 22 August 2020

Abstract: One of the important criteria used in judging the performance of a chatbot is the ability to provide meaningful and informative responses that correspond with the context of a user's utterance. Nowadays, the number of enterprises adopting and relying on task-oriented chatbots for profit is increasing. Dialog errors and inappropriate response to user queries by chatbots can result in huge cost implications. To achieve high performance, recent AI chatbot models are increasingly adopting the Transformer positional encoding and the attention-based architecture. While the transformer performs optimally in sequential generative chatbot models, recent studies has pointed out the occurrence of logical inconsistency and fuzzy error problems when the Transformer technique is adopted in retrieval-based chatbot models. Our investigation discovers that the encountered errors are caused by information losses. Therefore, in this paper, we address this problem by augmenting the Transformer-based retrieval chatbot architecture with a memory-based deep neural attention (mDNA) model by using an approach similar to late data fusion. The mDNA is a simple encoder-decoder neural architecture that comprises of bidirectional long short-term memory (Bi-LSTM), attention mechanism, and a memory for information retention in the encoder. In our experiments, we trained the model extensively on a large Ubuntu dialog corpus, and the results from recall evaluation scores show that the mDNA augmentation approach slightly outperforms selected state-of-the-art retrieval chatbot models. The results from the mDNA augmentation approach are quite impressive.

Keywords: Bi-LSTM; memory; NLP; attention; dialog-system; retrieval

1. Introduction

Many studies on natural language processing (NLP) agree that chatbots have contributed immensely towards the advancement in information retrieval and exchange between humans and computers. In the vertical domain, the profitability of online transactions in e-commerce, reservations, and marketing firms, etc., is beginning to depend heavily on chatbots to convince people to make

purchase of goods and services through interactive and persuasive chats. For this reason, it is important that a chatbot is able to respond coherently and accurately with respect to the context of queries from users. Therefore, we opine that the degree of the accuracy of selected response is the primary criteria to be used in judging the performance of a chatbot. However, modeling a chatbot to consistently select and match accurate responses with respect to the intent and context of the users' input utterances over multiple-turns of a conversation is a challenging task.

To address this challenge, a plethora of research studies have been done on retrieval-based dialog systems. These systems are more commonly adopted in vertical domains unlike in open domains where sequence-to-sequence generative chatbots are more prevalently adopted. Retrieval chatbots enjoy the advantage of possessing rich repositories from which they can select information-rich responses [1,2] that align within the scope of a predefined knowledge base. Whereas the generative-based ones possess so much liberty on how to respond to utterances, and thus, are usually prone to occasionally generate grammatical errors or irrelevant generic responses that may not serve the intents of the user.

Early studies on retrieval chatbots considered only data in single-turn of a dialog for selecting responses from a repository [3–5]. In-other-words, this approach entails sole use of the information in the last conversational utterance for matching a response. Meanwhile, outcomes from recent works such as the deep attention matching network (DAM) by Zhou et al. [1], shows that the consideration of information in multiple-turns of dialog prove to offer better performance in-which response selection matching not only considers information in last utterance turn, but also considers previous turns of utterances for matching contextual dependencies, which significantly improve accuracy of output response selection. This approach is based on the notion that human conversational manner often depend on the context and sentiments in the previous turns of utterances containing varying segments of semantic cues.

Recently, most of the state-of-the-art multi-turn retrieval chatbot models adopt the novel Transformer architecture proposed by Vaswani et al. [6] because an attention-based network helps neural systems to expedite the capturing of utterance-response textual dependencies by replacing deep tensors, thus minimising computational cost [7]. However, Transformer-based retrieval chatbot models [1,2,8] encounter the problematic occurrence of the following errors:

- **logical inconsistency error**: where retrieved response candidates are wrong due to logical mismatch
- **vague or fuzzy response candidate error**: where selected responses contains improper details

To contribute towards solving the above outlined problems, we propose the memory-based deep neural attention (mDNA). The key purpose of the mDNA is to augment transformer-based retrieval chatbot models to improve response accuracy and logical consistency. The mDNA is a deep neural network (DNN) architecture with a Bi-LSTM encoder integrated with attention mechanism; and in addition, a memory module is implemented in the encoder. The purpose of implementing memory is to capture and store important utterance word embedding so as to offer a cognitive-like ability to the model. In our experiments, we adopted the DAM [1] as a case study representing the transformer-based retrieval chatbot model. We combined the two model architectures (i.e., DAM and mDNA) with an approach similar to unimodal late fusion method [9], as shown in Figure 1, to achieve enhanced performance.

During the response retrieval process, the previous multi-turns of relevant utterance information that could have probably been "forgotten" (i.e., data lost in DAM but retained in mDNA memory) are parsed through the mDNA Bi-LSTM encoder with the attention mechanism to match long-range contextual dependencies. The final response candidates from the DAM and mDNA are combined to undergo a voting process via a softmax function for the final response output selection. Our experiments show that the fusion of two different model architectures increases the number of response candidates. It also increases the probability of the chatbot to select more logically consistent and accurate responses. Thus, this technique drastically minimises the problematic occurrence of fuzzy response candidates and logical response inconsistency errors impairing the performance of most transformer-based retrieval dialog systems.

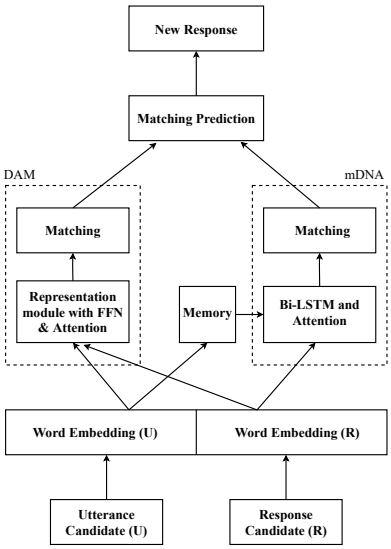

Figure 1. Operational flow of DAM augmented with mDNA.

2. Related Works

The growing number of research interests in multi-turn retrieval chatbots is an indication that it is a successful approach, but plagued with inherent challenges. Investigative studies done in [1,2,8] and so on, has sufficiently proven the superiority of multi-turn models over the single-turn ones. Therefore, with greater advantages, multi-turn retrieval chatbots are widely being adopted for various end-to-end task completion services [10] in the vertical domain which justifies their importance. Currently the existing efforts improve intuitive matching of utterance-context and semantic dependency to achieve high accuracy in response selection. It has driven recent scholarly works on retrieval chatbots [1,8,11,12] to adopt the use of Transformer attention-based architecture proposed in [6]. Although feed-forward neural networks (FFN) such as CNN are best suited for computer vision tasks, the Transformer attention model architecture has successfully used it for improving utterance textual context and response dependency matching in sequential neural dialog system. In addition, it has also achieved fair results in retrieval-based dialog system models such as the DAM network [1].

While the Transformer is now being hailed as the potential replacement to gated-recurrent neural networks in NLP tasks. Recent novel retrieval chatbot models based on the transformer-attention architecture [1,8,11] have shown to encounter the problematic occurrence of context mismatch and logical inconsistency errors in the selected responses of the models. Although some comparative studies on Transformer and RNN [13] agree that the self-attention and positional encoding attributes of the Transformer give them better performance leverage over the gate-based RNN neural models, more recent studies (e.g., [14]) point out that the multi-head attention mechanism of transformer models can also cause it to suffer loss of sequential information that are important in natural languages. Wang et al. [14] thereby proposed a RNN-enhanced Transformer (R-Transformer) model to address this problem.

However, we argue that by systematically applying memory to gated RNN-based models, it would achieve comparable long-range contextual matching performance such as the Transformer without suffering significant loss of important data. Scholarly works that have investigated the integration of memory in gate-based recurrent neural models include Zhao et al. [11] who used background document to supply external knowledge to retrieval chatbot models (i.e., a memory-like operation) to enrich response coherence with respect to the context of utterances. In other studies, memory support has been added to neural networks for natural language transduction tasks in [15]. Also, in [16], an attempt has been made to replace the use of attention with active memory. Furthermore, Wulamu et al. [17] combined the use of memory networks and attention to achieve

improved responses on simple Q&A tasks but, in more complex Q&A tasks, their model performed poorly. Owing to the poor model performance of [17] in handling complex retrieval tasks, the authors consider implementing Bi-LSTM and attention mechanism in their future work. Finally, investigating the effectiveness of augmenting Transformer with RNN-based neural model to forge a single robust ensemble, the research by Amazon's Domhan [18] suggests that one can successfully achieve good performance by such combination of architectures.

In this paper, we describe how we tackled the problems of logical mismatch and context inconsistency errors of multi-turn response selection in Transformer-based retrieval chatbots, by exploiting the collective benefits of memory, gated-RNN, and the Transformer attentional components combined as one unit. We selected the novel DAM network [1] to serve as a case study representing many other similar multi-turn retrieval chatbot models based on the Transformer architecture. By augmenting the DAM model with the mDNA, we achieved optimized performance, with respect to logical consistency and contextual accuracy of output responses.

3. Model Description

This section briefly describes the interactive roles of the major functional units that make up the mDNA (i.e., the Memory Encoding, Bi-LSTM with Attention, and the matching composition or prediction) as illustrated in Figure 2.

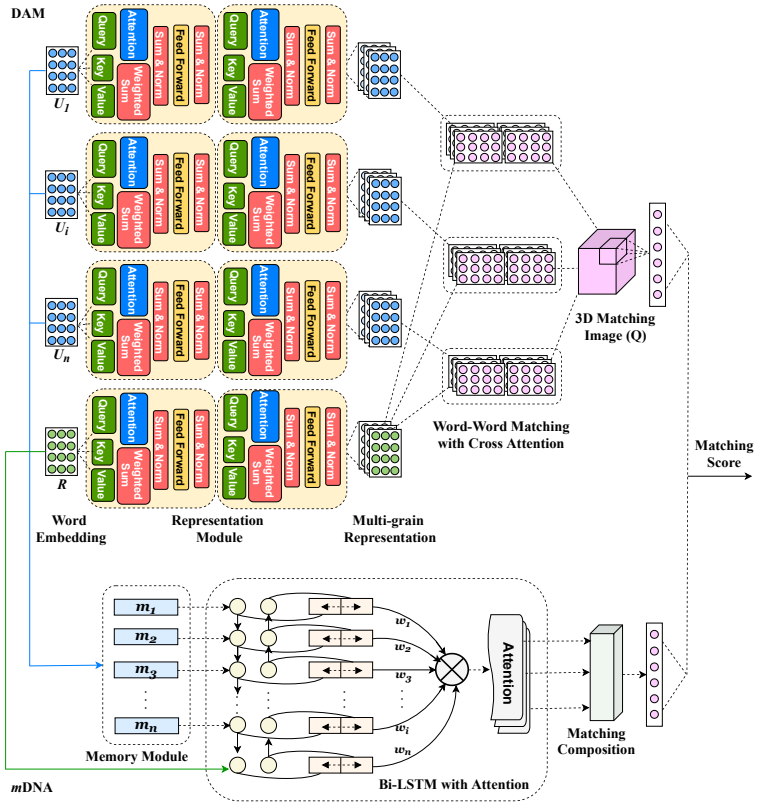

Figure 2. An overview of the combined architectures of DAM and mDNA.

3.1. Memory Utterance Embedding

The Transformer-based models can effectively capture and match textual dependencies in utterance and response. However, important relevant word representations in long multi-turn dialogs can easily be lost in the process and thereby impair the quality of the final predicted output responses. On this basis, we formulated the concept of the mDNA to comprise a memory unit for temporary

storage of the utterance input so as to retain important contextual utterance information which will be needed consequently for response dependency matching during a conversational session.

Here, we denote the input utterance embedding parsed from word2vec to the memory as $m = (m_1, m_2, m_3, \ldots, m_n)$. The response embedding is directly fed into the Bi-LSTM encoder and is denoted as $r = (r_1, \ldots, r_m)$. Then using the pre-trained word embedding table, we generate two input sequences as $M = [e_{m,1}, \ldots, e_{m,n}]$ and $R = [e_{r,1}, \ldots, e_{r,m}]$, where $e \in R^d$, and d is the dimension of word embedding. As shown in Figure 3, we use soft attention denoted as (α_M) to filter important utterance words vectors of given lengths e_ml for storage. We express this as, $\alpha_M = \text{softmax}(W, e_{m,l})$, where W is the weight score, and $e_{m,l}$ is utterance vector representation (m), and the length (l) of each word.

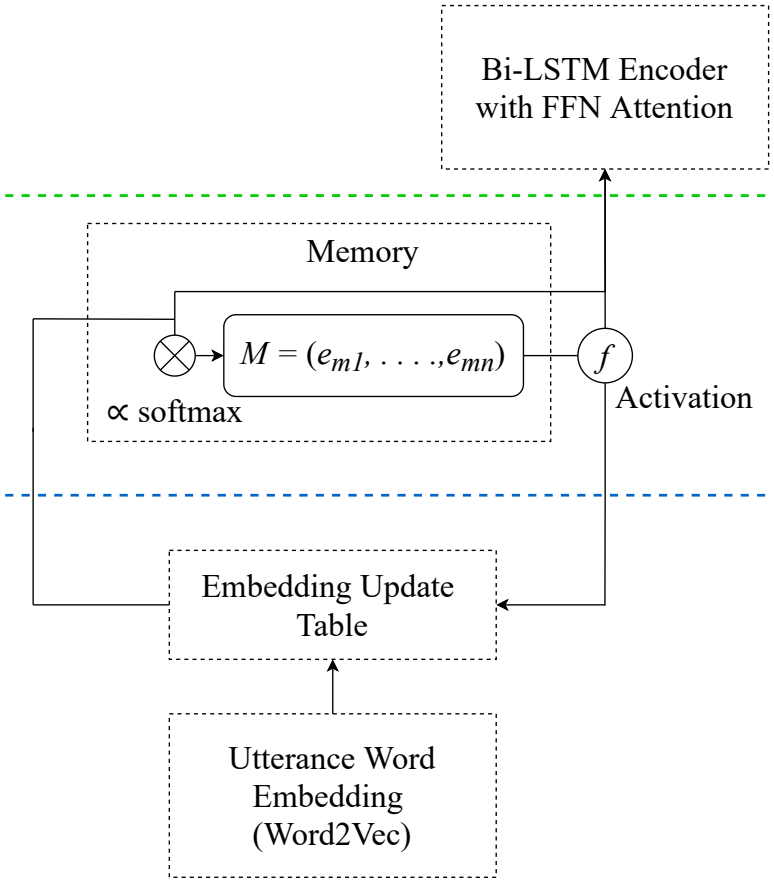

Figure 3. Storage system concept of the memory unit.

3.2. Bi-LSTM Encoding with Attention

Having solved the exploding and vanishing gradient problems of RNN, LSTM has enjoyed a lengthy time-line of being the basic model widely used in NLP tasks due to its capability to capture long-term textual dependencies [19]. Moreover, the enhanced bi-directional variant of the traditional LSTM which comprises of 2 hidden layers in opposing directions enables it to connect past and future input features to the same output in specified time steps, thus, the Bi-LSTM is able to learn fast. Meanwhile, comparing Bi-LSTM to the bi-directional gated recurrent unit (Bi-GRU) [20,21], the study [20] showed that the latter is equally fast; however, we justify our choice of adopting the Bi-LSTM in the mDNA model based on the analysis in [22] that the LSTM is a better option than GRU for performing tasks where deep context understanding is the major consideration of our study.

In this section, we apply the two input sequences of memory and response, M and R, into the Bi-LSTM encoder to represent the tokens of the contextual vectors, and we denote the vectors of the hidden state as m^s and r^s:

$$m_i^s = \text{BiLSTM}(M, i), \tag{1}$$

$$r_j^s = \text{BiLSTM}(R, j), \tag{2}$$

where i and j represent the i-th context in the utterance, and the j-th context in the response, respectively.

To determine whether a selected response candidate is appropriate and correlated with the context of the utterance, modeling the relationship between the utterance and the response is an essential step. For example, appropriate responses consider contextual keyword vectors, which can be obtained by modeling the semantic relationship. To this regard, being that not all words contribute equally to the context representation of an utterance, attention is required to ensure that important information are collected for matching composition. In this case, we use a multi-head attention mechanism to align the utterance contexts to the response candidates, and then calculate the semantic relationship at the utterance level. We also applied a soft alignment layer to calculate the attention weights as follows:

$$e_{ij} = (m_i^s)^T * r_j^s, \tag{3}$$

where $e \in R^{w*n}$, and T is the length of the sequence. For the hidden state of utterance context, i.e., m_i^s (already encoding the utterance and its contextual meaning), the relevant semantics in the response options are identified and composed using e_{ij} as a vector m_i^d, more specifically as shown in Equation (4).

$$\alpha_{ij} = \exp(e_{ij}) / \sum_{k=1}^{n} \exp(e_{ik}), m_i^d = \sum_{j=1}^{n} \alpha_{ij} r_j^s, \tag{4}$$

$$\beta_{ij} = \exp(e_{ij}) / \sum_{k=1}^{w} \exp(e_{kj}), r_j^d = \sum_{i=1}^{w} \beta_{ij} m_i^s, \tag{5}$$

where $\alpha \in R^w * n$ and $\beta \in R^w * n$ are the normalized attention weight matrices. The same process is also performed for each utterance and response matching in Equation (5).

Hence, we model the utterance level semantic relationship between the aligned utterance pairs, i.e., $< m_i^s, m_i^d >$ and $< r_i^s, r_i^d >$. The equation is further expressed as follows:

$$m_i^w = P([m_i^s; m_i^d; m_i^s - v_i^d; m_i^s \odot m_i^d]), \tag{6}$$

$$r_i^w = P([r_i^s; r_i^d; r_i^s - r_i^d; r_i^s \odot r_i^d]), \tag{7}$$

where a matching layer can be used to model some high-order interaction between the vectors m_i^w and r_i^w for the utterance and response, respectively. P is a 1-layer multi-layer perception (MLP) with a rectified linear unit (ReLU) activation.

3.3. Matching Composition and Prediction

In the process to predict the final response candidate selection, the Bi-LSTM is used to compose the matching vectors in a dense layer as follows:

$$m_i^m = \text{BiLSTM}(m^w, i), \tag{8}$$

and

$$r_j^m = \text{BiLSTM}(r^w, j). \tag{9}$$

Finally, we get the representations of DAM and mDNA denoted as $[f_1, f_2]$ to compute the final matching score $g(m, r)$, which is formulated as:

$$g(m, r) = \text{softmax}(W_2[f_1, f_2] + b_2), \tag{10}$$

where W_2 and b_2 are the learning parameters. The loss function of our model is the negative log likelihood, defined as:

$$L = -\sum_{D}[y \log(g(m,r)) + (1-y)(1-\log(g(m,r)))], \tag{11}$$

where D is the dataset, and y is the label marked in D.

4. Experiments

4.1. Dataset

The availability and use of a large amount of training data promises an increase in the probability of achieving good performance in neural machine translation tasks. For this reason, we used the Ubuntu Dialogue Corpus [23] to train and analyse the performance of our mDNA model. The Ubuntu Corpus is an English dataset which contains multi-turn dialogues constructed from Ubuntu Internet Relay Chat (IRC) logs. The dataset consists of one million utterance-response pairs, and each pair have a binary label to mark the responses as either positive or negative (as shown in Table 1). In the training set, the ratio of the positive and the negative is 1:1, and 1:9 in the validation and test sets. Table 2 gives the statistics of the training set, validation set, and test set.

Table 1. Sample of Ubuntu Dialogue output with Matching Score of mDNA.

		Label	**Matching Score**
Context	A: Hi, I am looking to see what packages are installed on my system, I don't see a path in the list being held somewhere else. B: try dpkg – get-selections. A: Is that like a database for packages instead of a flat file structure? B: dpkg is the debian package manager-get-selections that simply shows you what packages are handled by it.		
Response	No clue what do you need it for, its just reassurance as I don't know the debian package manager	1	0.995
	Then why not +q good point thanks.	0	0.231
	I mean real media ... not the command lol exactly.	0	0.035
	Hmm: should i force version to hoary.	0	0.299
	I will also run on core2 intels i installed ubuntu on a usb ...	0	0.016
	And what's your system specs the live cd does not use the hard drive at all	0	0.019
	Thanks i will see if i can find another option and use that as a last resort the steps ...	0	0.911
	Number is long term support also and not that much difference between number and number with the next ubuntu kubuntu and xubuntu at the end of april.	0	0.02
	These days backup to usb probably is more relevant ...	0	0.004
	It explains the same stuff a bit more in depth:) ...	0	0.963

Table 2. Statistics of Ubuntu Corpus dataset.

	Training	Validation	Testing
# context response pairs	1 M	500 K	500 K
# candidates per context	2	10	10
Avg. # turns per dialogue	7.71	7.33	7.54
Avg. # words per utterance	10.34	10.22	10.33

4.2. Hyperparameter Settings

We implemented the mDNA with Python 3 language, and used word2vec [24] to pre-train the word embedding on training set. The hidden size of the Bi-LSTM layers is set to 300, and to further enable fine-grain deep learning for good performance, the number of training epoch is set to 15 which we arrived at by fine-tuning the model with early stopping mechanism, while we set the mini-batch sizes to 16 for training, testing and validation, respectively, as could be accommodated by the Titan NVIDIA GPU memory resource available in our machine. Furthermore, the initial learning rate is set as 2×10^{-4} (i.e., 0.0002) used to update the learning parameters by stochastic gradient descent with Adam optimizer [25]. While low learning rate tends to slow down the training speed, it also contributes to steep decrease in the network's losses. Finally, the mDNA model is trained in Tensorflow-GPU environment with Cuda kernel, and the learning parameters W_2 and b_2 in the softmax function uses the default value set by Tensorflow. Being that Ubuntu dialog corpus is large (i.e., comprising of 1,000,000 Q&A pairs), we set the vocabulary size to 100,000; and lastly, we assigned the maximum lengths of utterance input embedding and response input embedding as 350 and 150, respectively.

5. Results and Evaluation

In this paper, we used the recall R at position k in n number of candidates, denoted as $R_n@k$, as the evaluation metrics, as it is also used in DAM [1], SMN [2], and other novel retrieval chatbot models [20,24,26–29] we selected to compare their performance with that of the mDNA model. Here, $R_n@k$ is defined as (the number of relevant retrieved response candidates at top-k)/(total n number of retrieved response candidates). We represent this mathematically as below:

$$R_n@k = \frac{TP}{TP + FN}, \tag{12}$$

where TP and FN represent true positive and false negative of retrieved response candidates, respectively. This evaluation metrics formula is suitable for information retrieval tasks, and is popularly used to measure and score the performance of information retrieval systems in terms of the degree at which the selected output response candidates match with the context of the input utterances. In our task, the mDNA model needs to select a response from n number of candidate responses. In this process, if the true response is among the response candidates, it is referred to as the true positive. Figure 4 shows the overview representation plot of $R_{10}@1$, 2, and 5, with respect to different values of epochs which shows that the model achieved a stable training (and learning). Moreover, to further analyze the performance of our model, we selected some representative models based on retrieval dialog systems (that are either based on Transformer or gated-RNN respective), DAM [1], SMN [2], DualEncoder [23], MV-LSTM [26], Match-LSTM [27], Multiview [28], DL2R [29], and BERT Bi-Encoder+CE [20]. We compared the performance of these models to that of the mDNA model; and to perform ablation analysis, we removed the DAM model combined with the mDNA, and referred to it as $mDNA_{self}$.

In Table 3, the recall evaluation results show that the mDNA performs competitively better than most of the selected chatbot models trained on Ubuntu Dialogue Corpus that were compared to it. With respect to the baseline DAM model, the performance of the mDNA increased by 3% on

R_{10}@1, 1.7% on R_{10}@2 and 0.6% on R_{10}@5, respectively. While the mDNA outperforms all other selected models compared to it, we observe that the recall evaluation scores of the mDNA and BERT Bi-Encoder+CE model [20] are similar, and therefore the performance gap between them is very insignificant. The BERT Bi-Encoder+CE model exploits the benefits of Bi-GRU and Bi-encoder system of BERT-Transformer architecture which enables its performance to be comparable to that of the mDNA. However, the capability of augmenting Transformer-based retrieval chatbot models for performance enhancement is a unique attribute to which the mDNA claims superiority.

Table 3. Performance evaluation results of the mDNA and some selected retrieval chatbot models trained on Ubuntu dataset.

Model	R_{10}@1	R_{10}@2	R_{10}@5
DualEncoder [23]	0.638	0.784	0.949
MV-LSTM [26]	0.653	0.804	0.946
Match-LSTM [27]	0.653	0.799	0.944
Multiview [28]	0.662	0.801	0.951
DL2R [29]	0.626	0.783	0.944
SMN [2]	0.726	0.847	0.961
DAM [1]	0.767	0.874	0.969
BERT Bi-Encoder + CE [20]	0.793	0.893	0.975
mDNA	**0.797**	**0.891**	**0.975**
mDNA$_{self}$	**0.788**	**0.885**	**0.971**

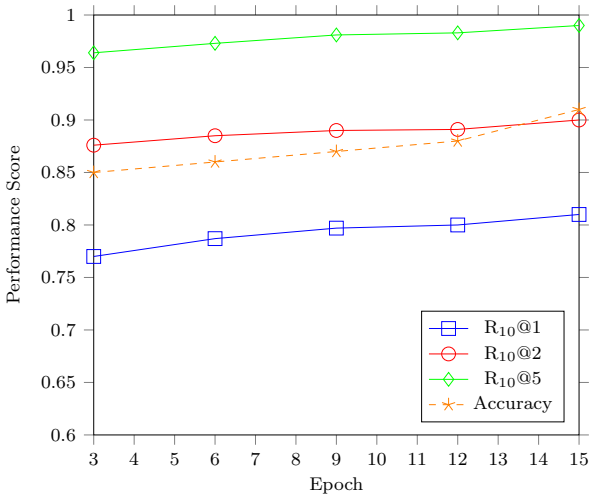

Figure 4. Plot showing increase in average accuracy (denoted as accuracy) scores and performance evaluation scores as the number of training epoch increases. The accuracy is averaged by R_{10}@1, R_{10}@2, and R_{10}@5 at corresponding epochs.

Lastly, owing to the important role of memory in the mDNA, we investigated to see if adjustment of the maximum length of utterance embedding is fed into the memory would affect the performance and effectiveness in any way. In this task, we initially set the dimension of utterance memory embedding to 350 and, subsequently, during multiple retraining sessions of the network, we altered the value from 350 to 300, 250, 200, and 150 at different corresponding intervals of the following training epochs: 3, 6, 9, 12, and 15, respectively. After several retraining of the model, we observed that the size of utterance word embedding parsed into the memory does not affect its performance, but instead, altering the values of the epochs directly influences overall accuracy and performance of the model. As the number of epoch increases from a lower value to the max set value, the average accuracy score also increases, as well as the scores of $Recall_{10}$@1, $Recall_{10}$@2, $Recall_{10}$@5, and vice-versa. This scenario is depicted with the plot in Figure 4.

6. Conclusions

In this paper, we presented the memory-based deep neural attention (mDNA) method and described the concepts of its operational process as well as the benefits of using fusion technique to augment independent retrieval chatbot models that are based on the Transformer architecture, for enhancing response selection accuracy, and consistency in logical context-dependency matching, respectively. As a case study, we selected the DAM network model and hybridised it with the mDNA to demonstrate the effectiveness of achieving better performance. The results obtained show the improvement in performance over some existing state-of-the-art retrieval chatbots. In future works, we will consider enhancing the Bi-LSTM encoder algorithm to support multimodal input-data and fusion (such as text and image) which will necessitate the use of more diversified datasets in the training of the model to further improve the accuracy of response selection.

Author Contributions: Conceptualization, J.C.; methodology, O.A.; writing–original draft preparation, J.C.; software, O.A.; validation, J.C. and O.A.; visualization, J.C.; supervision, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-182-042-MY2; and in part by the Chang Gung Memorial Hospital, Kweishan, Taoyuan, Taiwan, under Grants CMRPD2J0012 and CMRPD2I0052.

Conflicts of Interest: The authors declare no conflict of interests regarding the publication of this article.

References

1. Zhou, X.; Li, L.; Dong, D.; Liu, Y.; Chen, Y.; Zhao, W.X.; Yu, D.; Wu, H. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 1118–1127.
2. Wu, Y.; Wu, W.; Xing, C.; Li, Z.; Zhou, M. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 496–505.
3. Leuski, A.; Traum, D. NPCEditor: Creating Virtual Human Dialogue Using Information Retrieval Techniques. *AAAI AI Mag.* **2011**, *32*, 42–56. [CrossRef]
4. Wang, H.; Lu, Z.; Li, H.; Chen, E. A Dataset for Research on Short-Text Conversation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 935–945.
5. Hu, B.; Lu, Z.; Li, H.; Chen, Q. Convolutional neural network architectures for matching natural language sentences. In Proceedings of the 27th Conference Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2042–2050.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
7. Medina, J.R.; Kalita, J. Parallel Attention Mechanisms in Neural Machine Translation. In Proceedings of the 17th IEEE International Conference on Machine Learning and Applications, Orlando, FL, USA, 17–20 December 2018; pp. 547–552.
8. Zhang, Z.; Li, J.; Zhu, P.; Zhao, H.; Liu, G. Modeling Multi-turn Conversation with Deep Utterance Aggregation. In Proceedings of the 27th International Conference Computational Linguistics, Santa Fe, NM, USA, 21–25 November 2018; pp. 3740–3752.
9. Liu, K.; Li, Y.; Xu, N.; Natarajan, P. Learn to combine modalities in multimodal deep learning. *arXiv* **2018**, arXiv:1805.11730.
10. Henderson, M.; Vulic, I.; Gerz, D.; Casanueva, I.; Budzianowski, P.; Coope, S.; Spithourakis, G.; Wen, T.; Mrkšic, N.; Su, P. Training Neural Response Selection for Task-Oriented Dialogue Systems. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5392–5406.

11. Zhao, X.; Tao, C.; Wu, W.; Xu, C.; Zhao, D.; Yan, R. A Document-grounded Matching Network for Response Selection in Retrieval-based Chatbots. In Proceedings of the 28th IJCAI Conference on AI, Macao, China, 10–16 August 2019; pp. 5443–5449.
12. Yang, L.; Ai, Q.; Guo, J.; Croft, W.B. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In Proceedings of the 25th ACM International Conference Information and Knowledge, Indianapolis, IN, USA, 24–28 October 2016; Volume 8, pp. 287–296.
13. Lakew, S.M.; Cettolo, M.; Federico, M. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In Proceedings of the of 27th International Conference Computational Linguistics, Santa Fe, NM, USA, 20–25 August 2018; pp. 641–652.
14. Wang, Z.; Ma, Y.; Liu, Z.; Tang, J. R-transformer: Recurrent Neural Network Enhanced Transformer. *arXiv* **2019**, arXiv:1907.05572.
15. Sukhbaatar, S.; Szlam, A.; Weston, J.; Fergus, R. End-to-end Memory Networks. In Proceedings of the Conference on Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 1, pp. 2440–2448.
16. Kaiser, Ł.; Bengio, S. Can Active Memory Replace Attention? In Proceedings of the 30th Conference Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 3781–3789.
17. Wulamu, A.; Sun, Z.; Xie, Y.; Xu, C.; Yang, A. An Improved End-to-End Memory Network for QA Tasks. *Cmc Comput. Mater. Contin.* **2019**, *60*, 1283–1295. [CrossRef]
18. Domhan, T. How Much Attention Do You Need? A Granular Analysis of Neural Machine Translation Architectures. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1799–1808.
19. Zhou, Q.; Wu, H. NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification. In Proceedings of the 9th ACM Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Brussels, Belgium, 31 October 2018; pp. 189–194.
20. Vakili, A.; Shakery, A. Enriching Conversation Context in Retrieval-based Chatbots. *arXiv* **2019**, arXiv:1911.02290v1.
21. Chen, J.; Abdul, A. A Session-based Customer Preference Learning Method by Using the Gated Recurrent Units with Attention Function. *IEEE Access* **2019**, *7*, 17750–17759. [CrossRef]
22. Gruber, N.; Jockisch, A. Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text? *J. Front. Artif. Intell.* **2020**, *3*, 1–6. [CrossRef]
23. Lowe, R.; Pow, N.; Serban, I.; Pineau, J. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Proceedings of the of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czech Republic, 2–4 September 2015; pp. 285–294.
24. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th Annual Conference Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
25. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference Learning Representations, San Diego, CA, USA, 7–9 May 2015; Volume 1412.
26. Pang, L.; Lan, Y.; Guo, J.; Xu, J.; Wan, S.; Cheng, X. Text matching as image recognition. In Proceedings of the 30th AAAI Conference Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2793–2799.
27. Wang, S.; Jiang, J. Machine Comprehension using Match-LSTM and Answer Pointer. In Proceedings of the International Conference Learning Representations, Palais des Congrès Neptune, Toulon, France, 24–26 April 2017; pp. 2793–2799.
28. Zhou, X.; Dong, D.; Wu, H.; Zhao, S.; Yu, D.; Tian, H.; Liu, X.; Yan, R. Multi-view response selection for human-computer conversation. In Proceedings of the EMNLP 2016, Austin, TX, USA, 1–5 November 2016; pp. 372–381.
29. Yan, R.; Song, Y.; Wu, H. Learning to respond with deep neural networks for retrievalbased human-computer conversation system. In Proceedings of the SIGIR 2016, Pisa, Italy, 17–21 July 2016; pp. 55–64.