Extending the Transformer with Context and Multi-Dimensional Mechanism for Dialogue Response Generation

Ruxin Tan, Jiahui Sun, Bo Su(✉), and Gongshen Liu(✉)

School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai, China
{tanruxin,sjh_717,subo,lgshen}@sjtu.edu.cn

Abstract. The existing work of using generative model in multi-turn dialogue system is often based on RNN (Recurrent neural network) even though the Transformer structure has achieved great success in other fields of NLP. In the multi-turn conversation task, a response is produced according to both the source utterance and the utterances in the previous turn which are regarded as context utterances. However, vanilla Transformer processes utterances in isolation and hence cannot explicitly handle the differences between context utterances and source utterance. In addition, even the same word could have different meanings in different contexts as there are rich information within context utterance and source utterance in multi-turn conversation. Based on context and multi-dimensional attention mechanism, an end-to-end model, which is extended from vanilla Transformer, is proposed for response generation. With the context mechanism, information from the context utterance can flow to the source and hence jointly control response generation. Multi-dimensional attention mechanism enables our model to capture more context and source utterance information by 2D vectoring the attention weights. Experiments show that the proposed model outperforms other state-of-the-art models (+**35.8**% better than the best baseline).

Keywords: Multi-turn conversation · Response generation · Context and multi-dimensional mechanism.

1 Introduction

Generally, dialogue system is divided into two forms. One is task-oriented which is used to solve some specific problems [17], such as restaurant reservations, etc. The other is non-task-oriented, also called chatbot which is mainly used to chat with people [8]. Yan et al. reveal that non-task-oriented dialogue system on open domains is more common [19]. There are also two methods for non-task-oriented system: (1) Retrieval-based model, which learns to select the best response from the candidate repositories. (2) Generative model, which regards dialogue system problems as translation problems. As generative model can produce diverse responses, it is more challenging and attracts more attention.

Various generative models have been proposed to apply in dialogue system, such as sequence-to-sequence [11], hierarchical recurrent attention network [18]. Unfortunately, these models are all based on recurrent neural network (RNN) which needs to maintain chronological order. Therefore, the inherently sequential nature precludes parallelization. Recently, the Tranformer structure [16] has shown excellent performance especially in machine translation. But unlike the strong one-to-one correspondence between parallel language pairs, in dialogue system, there is often some dependency between adjacent utterances [13]. Zhou et al. have used the Transformer structure to solve the dialogue system problems but this structure is simply used in the retrieval-based model and only as an input layer to encode the input utterances [22]. More importantly, all those models mentioned above ignore the fact that different contexts and source utterance often have different effects on even the same word, and original attention cannot fully capture this difference [12].

In this work, we propose an end-to-end Transformer with context structure and multi-dimensional attention mechanism to solve the problems mentioned above. At the encoder side of the proposed model, an extra context encoder is added to handle context utterances. In the last layer of the encoder, we integrate the information of the context encoder into the source through the context-source attention mechanism and context gating sub-layer, which jointly affects the response generation of the decoder. By sharing the parameters with source encoder, the introduction of the context encoder does not add too many model parameters. At the same time, we introduce the multi-dimensional attention mechanism. Specifically, on the encoder side, the self-attention of the first layer in the Transformer is replaced by the multi-dimensional attention which calculates different attention weights for each feature of the word token, thus making full use of the alignment between each feature of the token and different utterances.

We have compared our proposed model with some state-of-the-art models using both automatic evaluation and side-by-side human comparison. The results show that the model significantly outperforms existing models on both metrics.

The key contributions of this paper can be summarized as follows:

(1) The proposed model is an end-to-end Transformer model for solving response generation in multi-turn conversation. It reveals that the Transformer model is effective in dialogue system field.

(2) Multi-dimensional attention mechanism is applied in dialogue system to better capture information within utterances.

2 Related Work

Shang et al. apply the basic recurrent neural network (RNN) encoder-decoder framework to handle the dialogue context and source [11]. Based on that, Serban et al. use hierarchical models (HRED) [9] and its variant (VHRED) [10] to capture the different impacts of utterances. Attention mechanism (HRAN) is then applied on these models [18]. They extend RNN to various hierarchical models to improve model performance at the expense of model complexity. The

Transformer is first used for machine translation [16], and Zhou et al. have applied it to dialogue but not an end-to-end training way [22].

Attention mechanism is widely used in NLP [1,16]. As the calculated attention score is a *scalar*, it does not take the effects of different contexts on each dimension of the word vector into account. To avoid it, Shen et al. propose multi-dimensional attention [12], the attention score is a 2D vector but so far there is no literature to prove whether the mechanism is applicable in dialogue system.

3 Approach

3.1 Problem Statement

Our goal is based on scenarios that generate responses in multi-turn conversation. Following Serban et al. [9], the data sets are triples $D = \{(U_{i,1}, U_{i,2}, U_{i,3})\}_{i=1}^{N}$, which represent *context utterance, source utterance, response utterance*, respectively. N is corpus size. $\forall i$, $U_{i,1} = (u_{i,1,1}, ..., u_{i,1,T_{i,1}})$, $U_{i,2} = (u_{i,2,1}, ..., u_{i,2,T_{i,2}})$, $U_{i,3} = (u_{i,3,1}, ..., u_{i,3,T_{i,3}})$ with their utterance length are $T_{i,1}, T_{i,2}, T_{i,3}$ respectively. Specifically, $u_{i,j,k}$ is the k-th word of the j-th utterance in one triple where the triple is the i-th within whole N size corpus. Similarly, $T_{i,j}$ is the length of the j-th utterance in one triple where the triple is i-th, that is, $U_{i,j}$.

We aim to estimate a generation probability $p(U_3|U_1, U_2)$ and the proposed model is able to produce a totally new response $U_3 = (u_{3,1}, ..., u_{3,T_3})$ according to the generation probability. In the next part, we will explain how to integrate context information into source and how to combine multi-dimensional attention.

3.2 RNN-Based Model and Transformer

In this part, RNN-Based Model used in dialogue system are introduced briefly.

RNN-Based Model Generally, on the encoder side, the RNN reads one word at one time step and stops until the utterance end flag is read. Then, the decoder starts decoding according to the state of the encoder and each time step decodes a word until the end flag. Details are as follows.

Given a context utterance $U_1 = (u_{1,1}, ..., u_{1,T_1})$[1], and source utterance $U_2 = (u_{2,1}, ..., u_{2,T_2})$, the encoder first calculates the hidden state of U_1:

$$h_{1,t} = f(u_{1,t}, h_{1,t-1}) \tag{1}$$

Where f is an RNN function unit such as LSTM [4] or GRU [2]. The last hidden state of the context utterance side h_{1,T_1} is used as the initial state of the source utterance side [11]:

$$h_{1,T_1} = h_{2,0} \tag{2}$$

[1] The subscript i is omitted for clarity

Similarly,

$$h_{2,t} = f(u_{2,t}, h_{2,t-1}) \tag{3}$$

Then the last hidden state of the source utterance h_{2,T_2} is used as the context vector c to produce response $U_3 = (u_{3,1}, ..., u_{3,T_3})$ word by word:

$$c = h_{2,T_2} \tag{4}$$
$$s_t = f(u_{3,t-1}, s_{t-1}, c) \tag{5}$$
$$p_t = softmax(s_t, u_{3,t-1}) \tag{6}$$

Where s_t is the hidden state of the decoder and p_t is the probability distribution of candidate words at time t. The context vector c can be calculated at different times for decoding. Specifically, each $u_{3,t}$ corresponds to a context vector c_t:

$$c_t = \sum_{j=1}^{T_1+T_2} \alpha_{t,j} h_j \tag{7}$$

$$h_j = \begin{cases} h_{1,j}, j < T_1 \\ h_{2,j-T_1}, j \geq T_1 \end{cases} \tag{8}$$

Where $\alpha_{t,j}$ is given by:

$$\alpha_{t,j} = \frac{exp(e_{t,j})}{\sum_{k=1}^{T_1+T_2} exp(e_{t,k})} \tag{9}$$

$$e_{t,j} = g(s_{t-1}, h_j) \tag{10}$$

Where g is a multilayer perceptron. Notice that $e_{t,j}$ and $\alpha_{t,j}$ are both *scalar*.

3.3 Our Model

Figure 1(b) shows our proposed model architecture. Compared with vanilla Transformer (see in Figure 1(a)), we keep the original decoder part unchanged and extend the encoder part with a context encoder to handle context utterance. In addition, we replace the attention mechanism used in the Transformer with the multi-dimensional attention mechanism.

Multi-Dimensional Attention The attention weights $e_{t,j}$ and $\alpha_{t,j}$ calculated by both RNN and the Transformer models are *scalar*, but multi-dimensional mechanism will calculate an attention value for each dimension of the word vector, thus producing a 2D vector attention weights [12]. For the original attention method as shown in Eq.10, the multilayer perceptron function of $\mathbf{g}$ is given by:

$$g(s_{t-1}, h_j) = w^T \sigma(W_1 h_j + W_2 s_{t-1}) \tag{11}$$

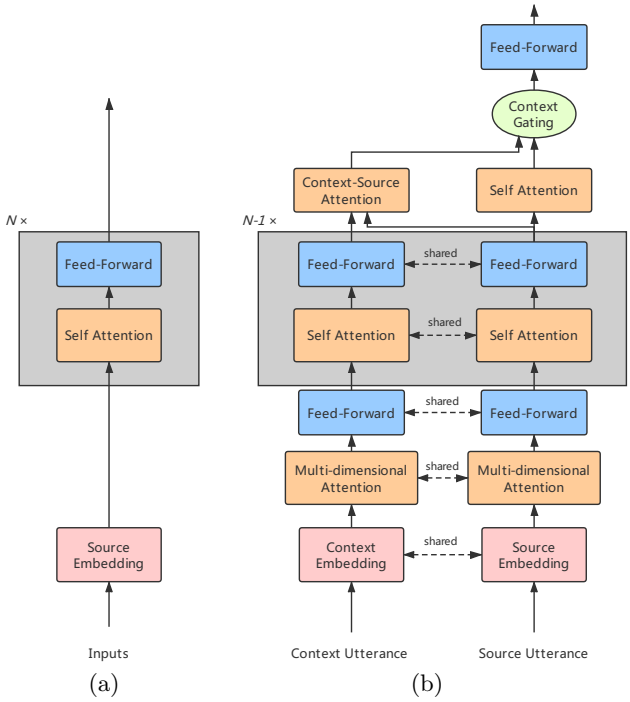

Fig. 1. (a) Vanilla Transformer encoder and (b) Our Transformer encoder with context and multi-dimensional mechanism

where we assume $h_j, s_{t-1} \in \mathbb{R}^d$, $w^T \in \mathbb{R}^d$ and d is the dimension of model. $W_1, W_2 \in \mathbb{R}^{d \times d}$ are parameters to learn. So the output of $\mathbf{g}$, i.e. attention weights are *scalar*. In multi-dimensional attention, the Eq.11 is replaced with:

$$g(s_{t-1}, h_j) = W^T \sigma(W_1 h_j + W_2 s_{t-1}) \tag{12}$$

$W^T \in \mathbb{R}^{d \times d}$ is a matrix. In this way, the attention weights distributed in each dimension of word vector can be obtained. The process is shown in Figure 2.

Transformer with Context Mechanism Context Encoder: At the bottom of our model, multi-dimensional attention is used to directly handle the word embedding of context utterance and then we stack $N-1$ layers where each single layer contains two same sublayers: self-attention layer and feed-forward layer.

Source Encoder: The first N layers of the encoder are the same as the context encoder. In order to avoid excessive increase of model parameters, shared parameters source encoder is used. The key issue is how to integrate the context utterance information into the source. Inspired by the idea of the encoder-decoder attention on the decoder side and the residual gating layer [6], in the last layer of the source encoder, we first use context-source attention to integrate

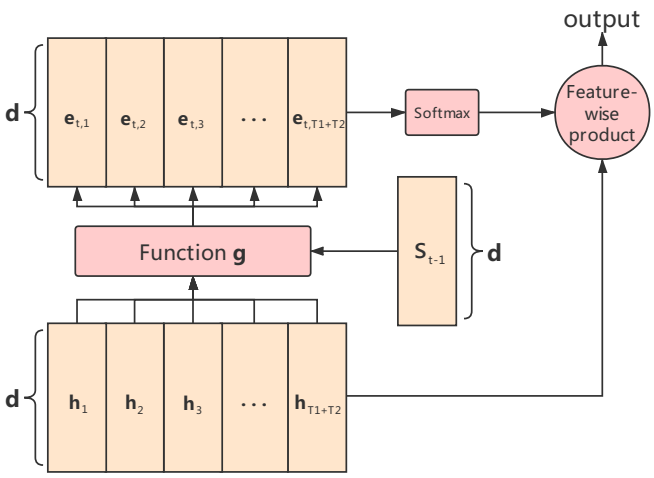

Fig. 2. Multi-Dimensional attention mechanism

context into source, then a *gate* is used to control the two attention mechanism information ratio between the output of self-attention on source and the output of context-source attention, which is

$$G = \sigma(W_G[C^{s-att}, C^{c-att}] + b_G) \tag{13}$$

$$C = G \odot C^{s-att} + (1 - G) \odot C^{c-att} \tag{14}$$

Where C^{s-att} is the output of source self-attention, C^{c-att} is the output of context-source attention with W_G, b_G being learning parameters. G is a function about the concatenation of C^{c-att} and C^{s-att}, C is their gated sum.

4 Experiments

4.1 Data Sets

The data set of dialogue system is a big challenge but not the issue of interest in this paper, so we use Douban [18,22], a commonly used data set, for evaluation.

To preprocess, we ensure that the length of each utterance is no more than 50 words. Following Serban et al. [9] and Tian et al. [15], we take the same way to divide the corpus into the triples $\{U1, U2, U3\}$. $U1$ stands for context, $U2$ stands for source and $U3$ stands for response. Thus, for Douban, there are $214,786$ training data, $1,682$ valid data, $13,796$ test data. We use context and source utterance to jointly generate source vocabulary and response utterance to generate target vocabulary, with each vocabulary size is $40,000$, covering 98.70% of words in context and source utterance, 99.06% of words in response utterance.

Table 1. BLEU scores, response length and entropy on different models

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Length	Entropy
S2S	3.1673	1.3067	0.9547	0.6935	5.2573	6.8027
HRED	4.5128	2.1032	0.6604	0.2215	6.6455	6.2849
VHRED	4.4365	1.8764	0.5314	0.1616	6.3933	6.3915
HRAN	4.6247	1.0947	0.7194	0.4529	5.2308	6.0189
Our Model	3.9658	1.4360	1.1581	**0.9419**	5.4480	**7.6751**

4.2 Baselines and Evaluation Metric

We use the following models as the baselines: (1) S2S (seq2seq) [11]. (2) HRED [9]. (3) VHRED [10]. (4) HRAN [18].

The parameters of all baseline models are subject to the settings in the original paper unless otherwise specified. All models are trained on a single GTX 1080 Ti GPU and implemented by the open source framework THUMT [21].

How to evaluate the performance of the generative model in dialogue system is also an open problem at present. Following Tian et al. [15], BLEU scores are selected as the evaluation metric which can measure the performance of generative model to some extent [14].

In addition, We use side-by-side human evaluation to verify performance between models and three volunteers are recruited[2] to manually score responses (see the paper [18] for more detail about how to conduct human evaluation).

4.3 Results and Analysis

For the proposed model, the number of encoder and decoder layers is $N = 6$, word embedding size is 512, model hidden size is $2,048$ and the number of head in self-attention is 8. During training, Adam [5] is used for optimization. In decoding, the beam size is set to 4.

Comparison with Baselines BLEU scores on different models are shown in Table 1. We can observe that our proposed model outperforms other models by a large margin under the more commonly used n-garm=4, that is, BLEU-4 with approximately **35%** higher than the second best model seq2seq (S2S). A similar situation exists in BLEU-3 as well.

Surprisingly, the performance improvement of the HRED and VHRED models is significant in BLEU-1 and BLEU-2. We further analyze the true output of different models and observe that HRED and VHRED are more inclined to output abundant repetitions of words like *the* or *you*, such responses are longer but universal and meaningless. Thus, the two models perform better when the n-gram value is smaller. We compute their response length and entropy [7]. The results are shown in Table 1. Our model has the highest entropy which means

[2] They are all native speakers and have graduate degrees or above.

Table 2. Human judgment results, *Win* means our model is better than the other model. We average the results of the three volunteers

Model	Win	Loss	Tie
Our Model v.s. S2S	**32.6**%	18.2%	49.2%
Our Model v.s. HRED	**38.9**%	13.8%	47.3%
Our Model v.s. VHRED	**38.2**%	13.5%	48.3%
Our Model v.s. HRAN	**33.7**%	15.6%	50.7%

Table 3. Ablation study results. CM and MDM denote the *context mechanism* and *multi-dimensional mechanism*. For vanilla Transformer, the input is a concatenation of context utterance and source utterance. Here BLEU means BLEU-4

Model	BLEU
Our Model	**0.9419**
No MDM	0.7999
No CM (vanilla Transformer)	0.6534

that the output diversity is higher. Even though the HRED and VHRED models output longer response, their entropy is much lower, it means that they are inclined to produce *safe* response.

Table 2 shows human judgment results compared with different models. Our model surpasses (win-loss) other models by 25.1% (HRED), 24.7% (VHRED), 18.1% (HRAN) and 14.4% (S2S), respectively.

Effect of Context Mechanism The experiment is conducted on the model without such mechanism, that is, vanilla Transformer. The results are shown in Table 3. Surprisingly, the result of vanilla Transformer is even slightly worse than seq2seq.

Further, we visually analyze the attention matrix in the context-source attention. Figure 3 is an example where context utterance is *Should (应该) be (是) viburnum (琼花)* and source utterance is *The emperor (隋炀帝) went down (下) Jiangnan (江南) just (就是) to (为了) see (看) it (它)*. As we expected, the most informative word of context utterance should be *viburnum* so it gets relatively larger attention than other words in context utterance. Also note the word *it* in source utterance, which correctly focuses more attention on the word *viburnum*.

Effect of Multi-Dimensional Mechanism The effect of multi-dimensional mechanism is shown in Table 3. By adding the multi-dimensional attention mechanism, our model performance is further improved by approximately **17%**.

We visualize the multi-dimensional mechanism. Figure 4 is the word "算账" multi-dimensional attention distribution heatmaps in two different utterances. One utterance is "我 算账 很好 谢谢" (*I am good at accounting, thanks.*), where "算账" here means *accounting*. The other is "这 喝酒 喝 得 好 难受 没 找 你

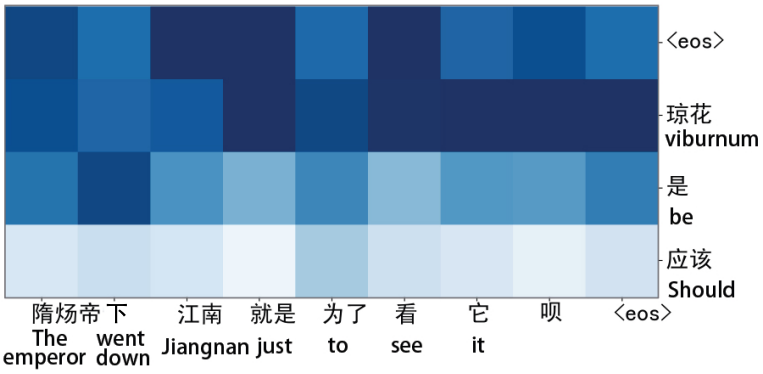

Fig. 3. A visual example of context-source attention. On the y-axis are the context tokens, on the x-axis are the source tokens.

Fig. 4. Multi-dimensional attention distribution comparison of the same word in different utterances. For the sake of illustration, we only show the first 100 dimensions in a 512D attention vector.

算账 不错 了 哈哈" (*This wine tastes terrible, you should be grateful that I did not find fault with you*), where "算账" means *find fault*. Within different utterances, nearly each dimension of multi-dimensional attention vector in the same word has totally different values. It indicates that the multi-dimensional attention mechanism is not redundant and even more important in a multi-turn conversation which needs more context-focused.

5 Conclusion

We extend the Transformer with context and multi-dimensional mechanism for multi-turn conversation. With our proposed mechanism, our model can better capture information from both context utterance and source utterance.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: EMNLP. pp. 1724–1734 (2014)
3. Gu, J., Bradbury, J., Xiong, C., Li, V.O., Socher, R.: Non-autoregressive neural machine translation. arXiv preprint arXiv:1711.02281 (2017)

4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Kuang, S., Xiong, D.: Fusing recency into neural machine translation with an inter-sentence gate model. arXiv preprint arXiv:1806.04466 (2018)
7. Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., Jin, Z.: Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In: COLING 2016: Technical Papers. pp. 3349–3358 (2016)
8. Ritter, A., Cherry, C., Dolan, W.B.: Data-driven response generation in social media. In: EMNLP. pp. 583–593. Association for Computational Linguistics (2011)
9. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI. vol. 16, pp. 3776–3784 (2016)
10. Serban, I.V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A.C., Bengio, Y.: A hierarchical latent variable encoder-decoder model for generating dialogues. In: AAAI. pp. 3295–3301 (2017)
11. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In: ACL (Volume 1: Long Papers). vol. 1, pp. 1577–1586 (2015)
12. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: Disan: Directional self-attention network for rnn/cnn-free language understanding. In: AAAI (2018)
13. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B.: A neural network approach to context-sensitive generation of conversational responses. In: NAACL: Human Language Technologies. pp. 196–205 (2015)
14. Tao, C., Mou, L., Zhao, D., Yan, R.: Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In: AAAI (2018)
15. Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., Zhao, D.: How to make context more useful? an empirical study on context-aware neural conversational models. In: ACL (Volume 2: Short Papers). vol. 2, pp. 231–236 (2017)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
17. Williams, J.D., Asadi, K., Zweig, G.: Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In: ACL (Volume 1: Long Papers). vol. 1, pp. 665–677 (2017)
18. Xing, C., Wu, Y., Wu, W., Huang, Y., Zhou, M.: Hierarchical recurrent attention network for response generation. In: AAAI (2018)
19. Yan, Z., Duan, N., Chen, P., Zhou, M., Zhou, J., Li, Z.: Building task-oriented dialogue systems for online shopping. In: AAAI. pp. 4618–4626 (2017)
20. Zhang, B., Xiong, D., Su, J.: Accelerating neural transformer via an average attention network. arXiv preprint arXiv:1805.00631 (2018)
21. Zhang, J., Ding, Y., Shen, S., Cheng, Y., Sun, M., Luan, H., Liu, Y.: Thumt: An open source toolkit for neural machine translation. arXiv preprint arXiv:1706.06415 (2017)
22. 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: ACL (Volume 1: Long Papers). vol. 1, pp. 1118–1127 (2018)