Constructing Interpretive Spatio-Temporal Features for Multi-Turn Responses Selection

Junyu Lu†, **Chenbin Zhang**†, **Zeying Xie, Guang Ling, Chao Zhou, Zenglin Xu**†
† SMILE Lab, University of Electronic Science and Technology of China, Sichuan, China
{cs.junyu, aleczhang13, swpdtz, zacharyling}@gmail.com,
tom.chaozhou@foxmail.com, zenglin@gmail.com

Abstract

Response selection plays an important role in fully automated dialogue systems. Given the dialogue context, the goal of response selection is to identify the best-matched next-utterance (i.e., response) from multiple candidates. Despite the efforts of many previous useful models, this task remains challenging due to the huge semantic gap and also the large size of candidate set. To address these issues, we propose a Spatio-Temporal Matching network (STM) for response selection. In detail, soft alignment is first used to obtain the local relevance between the context and the response. And then, we construct spatio-temporal features by aggregating attention images in time dimension and make use of 3D convolution and pooling operations to extract matching information. Evaluation on two large-scale multi-turn response selection tasks has demonstrated that our proposed model significantly outperforms the state-of-the-art model. Particularly, visualization analysis shows that the spatio-temporal features enables matching information in segment pairs and time sequences, and have good interpretability for multi-turn text matching.

1 Introduction

Fully automated dialogue systems (Litman and Silliman, 2004; Banchs and Li, 2012; Lowe et al., 2017; Zhou et al., 2018) are becoming increasingly important area in natural language processing. An important research topic in dialogue systems is response selection, as illustrated in Figure 1, which aims to select an optimal response from a pre-defined pool of potential responses (Kummerfeld et al., 2018). Practical methods to response selection are usually retrieval-based, that focus on matching the semantic similarity between the response and utterances in the dialogue history (Shang et al., 2015; Zhang et al., 2018).

Recently, convolutional operation, as a useful attempt to explore local correlation, has been investigated to extract the matching features from the attention grid (Wu et al., 2017; Zhou et al., 2018). Unfortunately, these methods usually do not perform well when there are many candidate responses.

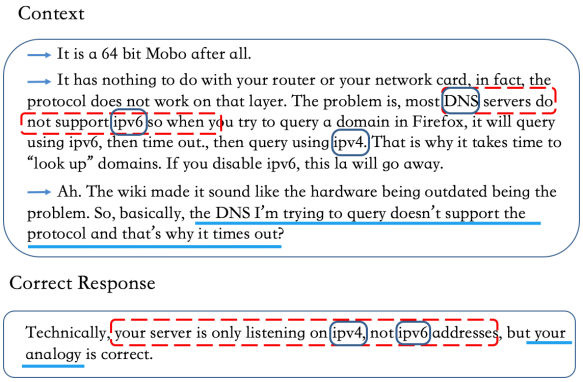

Figure 1: Examples of the Ubuntu dataset provided by **NOESIS** [1]. Text segments with the same color symbols across context and response can be seen as matched pairs.

In fact, in multi-turn dialogues, the next sentence is generally based on what was presented before and tends to match a recent local context. This is because the topic in a conversation may change over time, and the effective matching between the dialogue may only appear in a local time period. This phenomena generally appear in video processing (Hara et al., 2018; Tran et al., 2014), image caption (Chen et al., 2017) and action recognition (Girdhar and Ramanan, 2017).

Therefore, it is natural to adopt convolutional structure or attention mechanism to extract local matching information from the sentence sequences. Analogously, each turn of dialogue can be regarded as a frame of a video. This motivates us to propose the Spatio-Temporal Matching block (STM) to construct the spatio-temporal

[1]Noetic End-to-End Response Selection Challenge is described in detail at http://workshop.colips.org/dstc7.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 44–50
Florence, Italy, July 28 - August 2, 2019.

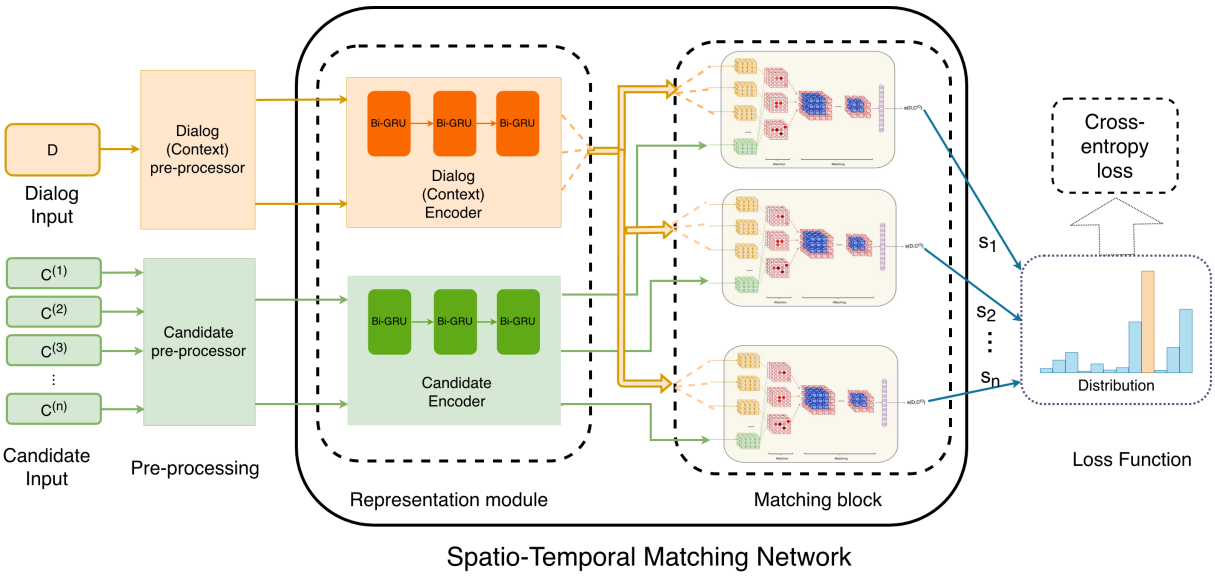

Figure 2: The proposed spatio-temporal matching framework for response selection.

features of local semantic relation between each turn of dialog and candidates by soft-attention mechanism. In detail, we model the response selection problem as a multi-class classification problem with sequences as input, where the label of the true response is set to one and the other candidates are set to zero. As illustrated in Figure 2, the proposed STM framework includes two parts: (i) **representation module** and (ii) **matching block**. Specifically, representations of the dialogue context and candidate answers are first learned through from dual encoders, and deep 3D ConvNets (Ji et al., 2013) are then used to match attentions between the dialogue contexts and candidate answers. Evaluation on the NOESIS datasets has demonstrated the outstanding performance of our proposed model against other well-known frameworks. Furthermore, our model enjoys a merit of good interpretation with the visualization of the attention weight as a thermal map. Our code is released under https://github.com/CSLujunyu/Spatio-Temporal-Matching-Network.

2 Our model

Before presenting the model, we first provide the problem formulation. Suppose that we have a dialogue dataset $\{(\mathbf{D}, \mathbf{C}, \mathbf{R})_i\}_{i=1}^{N}$, we denotes $\mathbf{D} = \{d_0, d_1, ..., d_m\}$ as a conversation context with utterances d_i and $\mathbf{C} = \{c_0, c_1, ..., c_n\}$ as the next utterance candidate set. $\mathbf{R}$ represents the correct response ID in the corresponding candidate set. Our goal is to learn a matching model between the dialog context $\mathbf{D}$ and the candidates c_i which can measure the matching degree and predict the best matched response.

2.1 Representation Module

Given a dialog context $\mathbf{D} = \{d_0, d_1, ..., d_m\}$ and candidates $\mathbf{C} = \{c_0, c_1, ..., c_n\}$, we employ L layers of bidirectional GRUs (Bi-GRU) (Cho et al., 2014) to extract sequential information in a sentence. The representations we used are deep, in the sense that they are a function of all of the internal layers of the Bi-GRU (Devlin et al., 2018; Peters et al., 2018a) We denote l^{th} GRU layer dialog and candidate representation as $\mathbf{H}_{\mu}^{l} = \{\mu_0^l, \mu_1^l, ..., \mu_m^l\}$ and $\mathbf{H}_{\gamma}^{l} = \{\gamma_0^l, \gamma_1^l, ..., \gamma_n^l\}$ respectively.

2.2 Spatio-Temporal Matching block

An illustration of the matching block is shown in Figure 3. We use attention mechanism to construct local related features for every candidate. In order to avoid the influence of gradient explosion caused by large dot product, matching matrices are constructed at each layer using scale-attention (Vaswani et al., 2017), which is defined as:

$$\mathbf{M}_{\mu_m,\gamma_n}^{l} = \frac{(\mu_m^l)^T \gamma_n^l}{\sqrt{d}}, \quad (1)$$

where $l \in [1, L]$, $\mu_m^l \in \mathbb{R}^{d \times n_\mu}$ denotes m^{th} turn of dialog representation at l^{th} GRU layer, $\gamma_n^l \in \mathbb{R}^{d \times n_\gamma}$ denotes n^{th} candidate representation at l^{th} GRU layer, $M_{\mu_m,\gamma_n}^{l} \in \mathbb{R}^{n_\mu \times n_\gamma}$ is constructed as

attention images, d is the dimension of word embedding, n_μ and n_γ denotes the number of words in dialog utterances and candidates respectively.

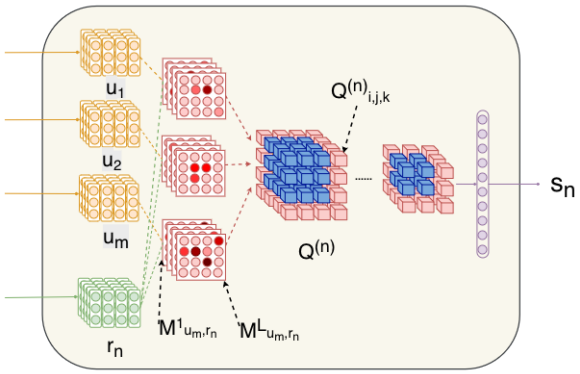

Figure 3: A close-up of the matching block

Moreover, in order to retain the natural temporal relationship of the matching matrices, we aggregate them all into a 4D-cube by expanding in time dimension. We call 4D-matching as spatio-temporal features and define images of n^{th} candidate as $\mathbf{Q}^{(n)}$:

$$\mathbf{Q^{(n)}} = \{Q^{(n)}_{i,j,k}\}_{m \times n_\mu \times n_\gamma}, \tag{2}$$

$$Q^{(n)}_{i,j,k} = \{M^l_{\mu_i,\gamma_n}[j,k]\}^L_{l=0}, \tag{3}$$

where $\mathbf{Q}^{(n)} \in \mathbb{R}^{m \times n_\mu \times n_\gamma \times L}$, $M^l_{\mu_i,\gamma_n}[j,k] \in \mathbb{R}$ and $Q^{(n)}_{i,j,k} \in \mathbb{R}^L$ is a pixel in $\mathbf{Q}^{(n)}$.

Motivated by C3D network (Tran et al., 2014), it is natural to apply a 3D ConvNet to extract local matching information from $\mathbf{Q}^{(n)}$. The operation of 3D convolution with max-pooling is the extension of typical 2D convolution, whose filters and strides are 3D cubes. Our matching block has four convolution layers and three pooling layers (First two convolution layers are both immediately followed by pooling layer, yet the last pooling layer follows two continuous convolution layers). All of 3D convolution filters are $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. With the intention of preserving the temporal information in the early phase, 3D pooling layers are set as $3 \times 3 \times 3$ with stride $3 \times 3 \times 3$ except for the first pooling layer which has kernel size of $1 \times 3 \times 3$ and stride $1 \times 3 \times 3$.

One fully-connected layer is used to predict the matching score between dialog context and potential responses. Finally, we compute softmax cross entropy loss,

$$s_n = \mathbf{W} f_{conv}(\mathbf{Q^{(n)}}) + \mathbf{b}, \tag{4}$$

where f_{conv} is the 3D ConvNet we used, $\mathbf{W}$ and $\mathbf{b}$ are learned parameters.

3 Experiments

3.1 Dataset

The ongoing DSTC series starts as an initiative to provide a common testbed for the task of Dialog State Tracking, and the most recent event, DSTC7 in 2018, mainly focused on end-to-end systems (Williams et al., 2013; Yoshino et al., 2019). We evaluate our model on two new datasets that released by the NOESIS (DSTC7 Track1): (1) **the Ubuntu Corpus**: Ubuntu IRC (Lowe et al., 2015a) consists of almost one million two-person conversations extracted from the Ubuntu chat logs , used to receive technical support for various Ubuntu-related problems. The newest version lies in manually annotations with a large set of candidates (Kummerfeld et al., 2018). The training data includes over 100,000 complete conversations, and the test data contains 1,000 partial conversations. (2) **the Advising Dataset**: It collects advisor dialogues for the purpose of guiding the student to pick courses that fit not only their curriculum, but also personal preferences about time, difficulty, career path, etc. It provides 100,000 partial conversations for training, obtained by cutting 500 conversations off randomly at different time points. Each conversation has a minimum of 3 turns and up to 100 candidates.

3.2 Metrics

We use the same evaluation metrics as in previous works and the recommendation of the NOESIS (Wu et al., 2017; Zhou et al., 2018; Yoshino et al., 2019). Each comparison model is asked to select k best-matched utterances from n available candidates. We calculate the recall of the true positive responses among the k selected ones and denote it as $\mathrm{R}_n@k = \frac{\sum_{i=0}^{k} y_i}{\sum_{i=0}^{n} y_i}$, where y_i is the binary label for each candidate. In addition, we use MRR (Mean reciprocal rank) (Voorhees et al., 1999; Radev et al., 2002) to evaluate the confident ranking of the candidates returned by our model.

3.3 Experimental Setting

We consider at most 9 turns and 50 words for each utterance and responses in our experiments. Word embeddings are initialized by GloVe[1](Pennington

[1]http://nlp.stanford.edu/data/glove.840B.300d.zip

Model	R_{100}@1	R_{100}@10	MRR
Baseline	0.083	0.359	-
DAM	0.347	0.663	0.356
DAM+Fine-tune	0.364	0.664	0.443
DME	0.383	0.725	0.498
DME-SMN	0.455	0.761	0.558
STM(Transform)	**0.490**	**0.764**	**0.588**
STM(GRU)	**0.503**	**0.783**	**0.597**
STM(Ensemble)	**0.521**	**0.797**	**0.616***
STM(BERT)	**0.548***	**0.827***	**0.614**

Table 1: Experiment Result on the Ubuntu Corpus.

Model	Advising 1		Advising 2	
	R_{100}@10	MRR	R_{100}@10	MRR
Baseline	0.296	-	-	-
DAM	0.603	0.312	0.374	0.174
DAM+Fine-tune	0.622	0.333	0.416	0.192
DME	0.420	0.215	0.304	0.142
DME-SMN	0.570	0.335	0.388	0.183
STM(Transform)	0.590	0.320	0.404	0.182
STM(GRU)	**0.654**	**0.380**	**0.466**	**0.220**
STM(Ensemble)	**0.662***	**0.385***	**0.502***	**0.232***

Table 2: Experiment Results on the Advising Dataset.

et al., 2014) and updated during training. We use Adam (Kingma and Ba, 2014) as the optimizer, set the initial learning rate is 0.001, and we employ early-stopping(Caruana et al., 2001) as a regularization strategy.

3.4 Comparison Methods

In this paper, we investigate the current state-of-the-art model in response selection task. In order to make it compatible to the task of NOESIS, we have made some changes as following: **(1) Baseline** The benchmark released by DSTC7 is an extension of the Dual LSTM Encoder model [2] (Lowe et al., 2015b). **(2) Dual Multi-turn Encoder** Different from Baseline, we use a multi-turn encoder to embed each utterance respectively and calculate utterance-candidate matching scores using dot product at the last hidden state of LSTM. **(3) Sequential Matching Network** We employ Sequential Matching Network (Wu et al., 2017) to measure the matching score of each candidate, and then calculate categorical cross entropy loss across all of them. We name it as DME-SMN in Table 1, 2. **(4) Deep Attention Matching Network** The DAM (Zhou et al., 2018) trained on undersampling data (Chawla, 2009), which use a 1:1 ratio between true responses and negative responses for training, is represented as DAM in Table 1, 2. Furthermore, we also construct context-related negative responses to train the model. We observe that using only this context-related negative responses to train the model will result in divergence. So this data is only used for fine-tuning. In this way, DAM is firstly trained on undersampling data then get fine-tuned with context-related negative responses. We name this model as DAM+Fine-tune in Table 1, 2.

[2] https://github.com/IBM/dstc7-noesis/tree/master/noesis-tf

3.5 Ablation Study

As it is shown in Table 1, we conduct an ablation study on the testset of the Ubuntu Corpus, where we aim to examine the effect of each part in our proposed model.

Firstly, we verify the effectiveness of dual multi-turn encoder by comparing Baseline and DME in Table 1. Thanks to dual multi-turn encoder, DME achieves 0.725 at R_{100}@10 which is 0.366 better than the Baseline (Lowe et al., 2015b).

Secondly, we study the ability of representation module by testing LSTM, GRU and Transformer with the default hyperparameter in Tensorflow. We note that GRU is better for this task. After removing spatio-temporal matching block, the performance degrades significantly.

In order to verify the effectiveness of STM block further, we design a DME-SMN which uses 2D convolution for extracting spatial attention information and employ GRU for modeling temporal information. The STM block makes a 10.54% improvement at R_{100}@1.

Next, we replace GRU with Transformer in STM. Supposed the data has maximal m turns and n candidates, the time complexity of cross-attention (Zhou et al., 2018), $O(mn)$, is much higher than that of the Dual-Encoder based model, $O(m+n)$. Thus, cross-attention is an impractical operation when the candidate set is large. So we remove cross-attention operations in DAM and extend it with Dual-Encoder architecture. The result in Table 1 shows that using self-attention only may not be enough for representation.

As BERT (Devlin et al., 2018) has been shown to be a powerful feature extractor for various tasks, we employ BERT as a feature-based approach to generate ELMo-like pre-trained contextual representations (Peters et al., 2018b).It succeed the

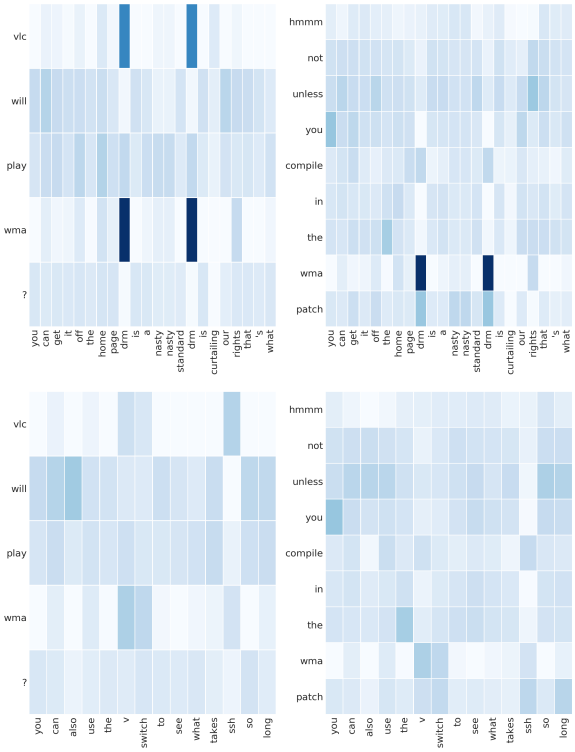

Figure 4: Attention feature across positive and negative matching in the first layer.

highest results and outperforms other methods by a significant margin.

3.6 Visualization

In order to demonstrate the effectiveness of spatio-temporal information matching mechanism, we visualize attention features across positive and negative examples.

To clarify how our model identifies important matching information between context and candidates, we visualize the attention matching matrices in Figure 4. The first row is positive matching matrices and the sencond is negative matching example. We denote the y-axis of Figure 4 as response sentence and the x-axis as utterances in context. Each colored grid represents the matching degree or attention score between two words. Deeper color represents better matching. Attention images in the first row are related to positive matching while those of the second row are related to negative matching. Intuitively, We can see that important words such as "vlc", "wma" are recognized and carried to match "drm" in correct response. In contrast, the incorrect response has no correlation and thus little matching spaces.

Note that our model can not only match word-level information, but also can match segment-level or sentence level information using 3D convolution. As it shows in Figure 5, the second layer tends to concentrate on segment-level information for which "wma patch" in utterance highly match "the home page drm" and "nasty nasty standard drm" in response. Furthermore, we find in our experiment that third layer tends to focus on sentence topic and more abstract meaning of the segments, which achieve better performance. However, more than three layers will destroy model ability in our experiments.

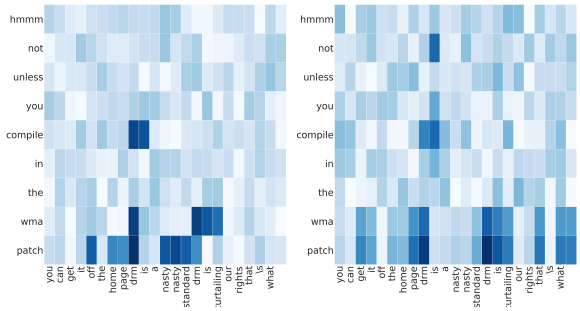

Figure 5: Attention feature in different granularities. Left picture represents the second layer matching matrix for segment granularities, while right picture match at the third layer.

4 Conclusion and Future Work

In this paper, we proposed an End-to-End spatio-temporal matching model for response selection. The model uses a dual stacked GRU or pre-trained BERT to embed utterances and candidates respectively and apply spatio-temporal matching block to measure the matching degree of a pair of context and candidate. Visualization of attention layers illustrates that our model has the good interpretative ability, and has the ability to pick out important words and sentences.

In the future, we would like to explore the effectiveness of various attention methods to solve indefinite choices task with interpretive features.

5 Acknowledgement

Junyu Lu, Chenbin Zhang and Zenglin Xu was partially supported by a grant from National Natural Science Foudation of China (No.61572111), Startup fundings of UESTC (Nos.A1098531023601041 and G05QNQR004), and a Research Fund for the Central Universities of China (No.ZYGX2016Z003).

References

Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Rich Caruana, Steve Lawrence, and C Lee Giles. 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. pages 402–408.

Nitesh V Chawla. 2009. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. pages 5659–5667.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rohit Girdhar and Deva Ramanan. 2017. Attentional pooling for action recognition. *Neural Information Processing Systems (NIPS)*, pages 34–45.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? pages 6546–6555.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jonathan K Kummerfeld, Sai R Gouravajhala, Joseph Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros Polymenakos, and Walter S Lasecki. 2018. Analyzing assumptions in conversation disentanglement research through the lens of a new dataset and model. *arXiv preprint arXiv:1810.11118*.

Diane J Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015a. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909.

Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. 2015b. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *Proceedings of the SIGDIAL 2015 Conference*, page 285294.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2014. C3D: generic features for video analysis. *CoRR*, abs/1412.0767.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413.

Yu Wu, Wei Wu, Chen Xing, Zhoujun Li, and Ming Zhou. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *Proceedings ofthe 55th Annual Meeting ofthe Association for Computational Linguistics*, pages 496–505.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, R. Chulaka Gunasekara, Walter S. Lasecki, Jonathan K. Kummerfeld, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Xiang Gao, Huda AlAmri, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2019.

Dialog system technology challenge 7. *CoRR*, abs/1901.03461.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018. Addressee and response selection in multi-party conversations with speaker interaction rnns.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. *Proceedings ofthe 56th Annual Meeting ofthe Association for Computational Linguistics*, pages 1–10.