Reading Turn by Turn: Hierarchical Attention Architecture for Spoken Dialogue Comprehension

Zhengyuan Liu, Nancy F. Chen
Institute for Infocomm Research, A*STAR
{liu_zhengyuan, nfychen}@i2r.a-star.edu.sg

Abstract

Comprehending multi-turn spoken conversations is an emerging research area, presenting challenges different from reading comprehension of passages due to the interactive nature of information exchange from at least two speakers. Unlike passages, where sentences are often the default semantic modeling unit, in multi-turn conversations, a *turn* is a topically coherent unit embodied with immediately relevant context, making it a linguistically intuitive segment for computationally modeling verbal interactions. Therefore, in this work, we propose a hierarchical attention neural network architecture, combining turn-level and word-level attention mechanisms, to improve spoken dialogue comprehension performance. Experiments are conducted on a multi-turn conversation dataset, where nurses inquire and discuss symptom information with patients. We empirically show that the proposed approach outperforms standard attention baselines, achieves more efficient learning outcomes, and is more robust to lengthy and out-of-distribution test samples.

1 Introduction

Reading comprehension has attracted much interest in the past couple years, fueled by avid neural modeling investigations. Given a certain textual content, the goal is to answer a series of questions based on implicit semantic understanding. Previous work has focused on passages like Wikipedia (Rajpurkar et al., 2016) or news articles (Hermann et al., 2015). Recently, dialogue comprehension in the form of cloze tests and multi-choice questions has also started to spur research interest (Ma et al., 2018; Sun et al., 2019). Different from passages, human-to-human dialogues are a dynamic and interactive flow of information exchange, which are often informal, verbose and repetitive.[1] This leads to lower information density and more topic diffusion, since the spoken content of a conversation is determined by two speakers, each with his/her own thought process and potentially distracting and parallel streams of thoughts.

To address such challenges, we propose to utilize a hierarchical attention mechanism for dialogue comprehension, which has shown to be effective in various natural language processing tasks (Yang et al., 2016; Choi et al., 2017; Hsu et al., 2018). The hierarchical models successively capture contextual information at different levels of granularity, leveraging coarse-grained attention to reduce the potential distraction in finer-grained attention but at the same time exploit finer-grained attention to distill key information for downstream tasks more precisely and efficiently.

While in document tasks sentences are the default semantic modeling unit at the coarse-grained level, utterances might be a more suitable counterpart in spoken dialogues, as dialogues often consist of incomplete sentences. However, a single utterance/sentence which usually implies information from one speaker is insufficient for grasping the full relevant context, as the interactive information from the interlocutor is often necessary. In multi-turn dialogues, each *turn* is one round of information exchange between speakers, thus making it a linguistically intuitive segment for modeling verbal communications. Thus, we postulate that for spoken dialogue comprehension, it is more effective to model conversations turn by turn using a multi-granularity design.

In this work, we introduce a hierarchical neu-

[1] One needs to process information on the spot during conversations, hence a particular concept could take rounds of interactions to confirm the information is conveyed correctly before moving on to the next topic, while for passages the reader can process the information at his own pace.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5460–5466
Florence, Italy, July 28 - August 2, 2019.

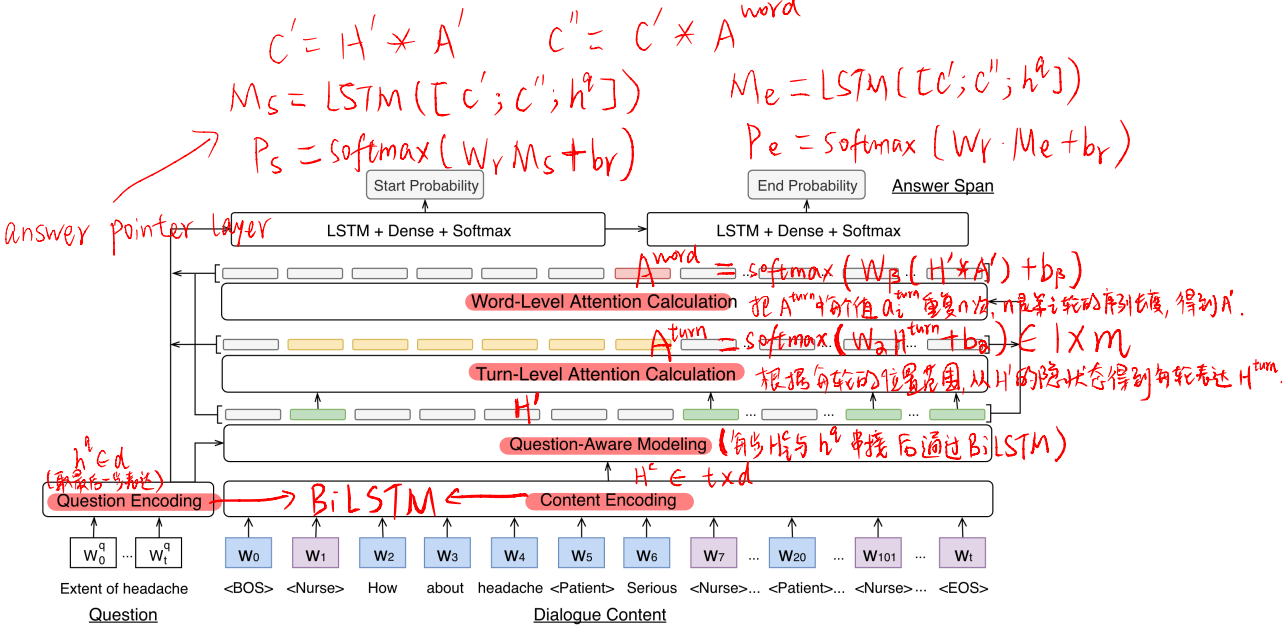

Figure 1: Turn-based hierarchical architecture for dialogue comprehension: tokens in purple are the indicators of dialogue turns, and their indices are used to select question-aware hidden states (Green) for turn-level attention calculation. The turn with higher attentive score (Yellow) contributes more in scoring word-level attentions (Red).

ral attention architecture, integrating turn-level attention with word-level attention for multi-turn dialogue comprehension in a question-answering manner, where we evaluate performance on a corpus preserving linguistic features from real-world spoken conversation scenarios. In particular, we examine how our approach is able to address challenges from limited training data scenarios and from lengthy and out-of-distribution test samples.

2 Hierarchical Attention Architecture

The proposed architecture of modeling multi-level attention for dialogue comprehension is shown in Figure 1. The model design is based on extractive question answering, and consists of the following layers: a sequence encoding layer, a question-aware modeling layer, a turn-level attention layer, a word-level attention layer, and an answer pointer layer. We elaborate on the details below.

2.1 Sequence Encoding Layer

Given a t-length sequence of word embedding vectors $S = \{w_0, w_1, ...w_t\}$, a bi-directional long short-term memory (Bi-LSTM) layer (Schuster and Paliwal, 1997) is used to encode S to a hidden representation, $H = \{h_0, h_1, ...h_t\} \in R^{t \times d}$, where d is the hidden dimension. We obtain the content representation H^c by encoding the dialogue sequence and concatenating the forward and backward information:

$$h_i^c = [LSTM_{w_i^c}^{Forward}; LSTM_{w_i^c}^{Backward}] \quad (1)$$

and extracting the last hidden state of question encoding as the question representation h^q.

2.2 Question-Aware Modeling Layer

We concatenate each step of H^c with the question h^q as in aspect-modeling (Wang et al., 2016), then obtain the question-aware modeling $H^{'}$ via a Bi-LSTM layer.

$$h_i^{'} = [LSTM_{[h_i^c;h^q]}^{Forward}; LSTM_{[h_i^c;h^q]}^{Backward}] \quad (2)$$

2.3 Turn-Level Attention Layer

We design the turn-level attention to score the dialogue turns explicitly, so the more salient turns will obtain higher scores, which is similar to (Hsu et al., 2018). However, instead of calculating the sentence-level attention using a separate recurrent component, we directly obtain the turn representations H^{turn} by collecting hidden states from $H^{'}$ with the turn-level segment position indices $T^{turn} = \{t_0^{turn}, t_1^{turn}, ...t_m^{turn}\}$, where m is the turn number of the dialogue content. More specifically, in a two-party conversation, each continuous utterance span between the speakers will be labeled as in one turn segment, and $t_{i+1}^{turn} - t_i^{turn}$ is the length of the ith turn.

Then the turn-level attentive score is calculated via a dense layer and softmax normalization:

$$A^{turn} = \text{softmax}(W_\alpha H^{turn} + b_\alpha) \quad (3)$$

2.4 Word-Level Attention Layer

In our hierarchical architecture, to mitigate adverse effects of spurious word-level attention from words in less attended turns, we utilize turn-level salient scores to modulate word-level attentions. Thus, we broadcast each a_i^{turn} in A^{turn} with its turn length to obtain $A^{'}$ in dialogue length, and

then multiply H' with A' to obtain the contextual sequence C'. Then the word-level attention A^{word} is calculated on C', and multiplied with H' to obtain the contextual sequence C''.

$$A^{word} = \text{softmax}(W_\beta(H' * A') + b_\beta) \quad (4)$$

2.5 Answer Pointer Layer

Contextual sequences C', C'' and question h^q are concatenated together and fed to a LSTM modeling layer. Then a dense layer with softmax normalization is applied for answer span prediction (Wang and Jiang, 2016).

$$M_{s/e} = LSTM_{[C';C'';h^q]} \quad (5)$$

$$P_{s/e} = \text{softmax}(W_\gamma M_{s/e} + b_\gamma) \quad (6)$$

where each p_s/p_e indicates the probability of being the start/end position of the answer span.

2.6 Loss function

Cross-entropy loss function is used as the metric between the predicted label and the ground-truth distribution. The total loss $\mathcal{L}_{\text{total}}$ contains the loss from answer span (Wang and Jiang, 2016) and from turn-level attentive scoring similar to (Hsu et al., 2018), with a weight $\lambda \in [0, 1]$.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{span}} + \lambda \mathcal{L}_{\text{turn\_attn}} \quad (7)$$

3 Experiments

3.1 Corpus & Data Processing

Dialogue Dataset: We evaluated the proposed approach on a spoken dialogue comprehension dataset, consisting of nurse-to-patient symptom monitoring conversations. This corpus was inspired by real dialogues in the clinical setting where nurses inquire about symptoms of patients (Liu et al., 2019). Linguistic structures at the semantic, syntactic, discourse and pragmatic levels were abstracted from these conversations to construct templates for simulating multi-turn dialogues. The informal styles of expressions, including incomplete sentences, incorrect grammar and diffuse flow of topics were preserved.

A team of linguistically trained personnel refined, substantiated, and corrected the automatically simulated dialogues by enriching verbal expressions through different English speaking populations in Asia, Europe and the U.S., validating logical correctness through checking if the conversations were natural, reasonable and not disobeying common sense, and verifying the clinical content by consulting certified and registered nurses. These conversations cover 9 topics/symptoms (e.g. headache, cough). For each conversation, the average word number is 255 and the average turn number is 15.5.

[Nurse] Hi Mr.#name#, you were discharged on #date#. There are some questions i'd like to check with you.
[Patient] Ok, Ok.
[Nurse] For these days, do you have some swelling or not?
[Patient] Swelling? It comes and go, comes and go.
[Nurse] Comes and go ... I see .. **#repetition#**
[Nurse] ... When it started?
[Patient] When I was travelling in #city#, last month.
[Nurse] Oh, you travelled in #city#, how is that?
[Patient] Great, I love the sunshine there.
... ...
[Nurse] So right now you talk to me, do you have any swelling?
[Patient] Umm ... **#backchannel#**
[Nurse] Let me check, the last time you told me is at your foot.
[Patient] Oh, right, only a bit.
... ...
[Nurse] Still feel some chest pain or tightness?
[Patient] No, do not have, chest tightness also not yet.
... ...
[Nurse] Any giddiness or heartbeat very fast?
[Patient] Heartbeat very fast? Do not have-- **#interruption#**
[Nurse] Well ... Do you-- **#interruption#**
[Patient] and no giddiness, no, nothing.
[Nurse] Ok, you need check your heartrate everyday.
[Nurse] Do you know how to use the machine?
[Patient] Yes, yes, no problem.
[Nurse] No heartbeat abnormal?
[Nurse] **#pause#** Because i see your heartbeat is a bit fast.
[Patient] Eighty five time, seems normal.
... ...

Figure 2: Examples of segmented turns in our corpus. The default segmented turn is an adjacency pair of utterances from two speakers (Yellow). To ensure a turn spans across semantically congruent utterances, neighboring utterances could be merged according to a set of rules derived from spoken features, like n-gram repetition (Green), back-channeling (Blue), self-pause (Red) and interlocutor interruption (Gray).

Turn Segmentation: In a smooth conversation, one turn is an adjacency pair of two utterances from two speakers (Sacks et al., 1974). However, in real scenarios, the conversation flow is often disrupted by verbal distractions such as interlocutor interruption, back-channeling, self-pause and repetition (Schlangen, 2006). We thus annotated these verbal features from transcripts of the real-world dialogues and integrated them in the templates, which are used to generate the simulated dialogue data. We subsequently merged the adjacent utterances from speakers considering the features and the intents to form turns (see Figure 2). This procedure ensures semantic congruence of each turn. Then the segment indices of turns were labeled for turn-level context collection.

Annotations for Question Answering: For the

comprehension task, questions were raised to query different attributes of a specified symptom; e.g., *How frequently did you experience headaches?* Answer spans in the dialogues were labeled with start and end indices, and turns containing the answer span were annotated for turn-level attention training.

3.2 Baseline Models

We implemented the proposed turn-based hierarchical attention (HA) model, and compared it with several baselines:

Pointer LSTM: We implemented a Pointer network for QA (Vinyals et al., 2015). The content and question embedding are concatenated and fed to a two-layer Bi-LSTM, then the answer span is predicted as in Section 2.5.

Bi-DAF: We implemented the Bi-Directional Attention Flow network (Seo et al., 2017) as an established baseline, which fuses question-aware and context-aware attention.

R-Net: We implemented R-Net (Wang et al., 2017), another established baseline, which introduces self-attention to implicitly model multi-level contextual information.

Utterance-based HA: To evaluate the effectiveness of turn-level modeling, we implemented an utterance-based model as the control, by treating every utterance as a single segment.

3.3 Training Configuration

Pre-trained word embeddings from Glove (Pennington et al., 2014) were utilized and fixed during training. Out-of-vocabulary words were replaced with the *[unk]* token. The hidden size and embedding dimension were set to 300. Adam optimizer (Kingma and Ba, 2015) was used with batch size of 64 and learning rate of 0.001. For the modeling layers, dropout rate was set to 0.2 (Srivastava et al., 2014). The weight λ in the loss function was set to 1.0. During training, the validation-based early stop strategy was applied. During prediction, we selected answer spans using the maximum product of p_s and p_e, with a constraint such that $0 \leq e - s \leq 10$.

3.4 Evaluation: Comparison with Baselines

Evaluation was conducted on the dialogue corpus described in Section 3.1, where the training, validation and test sets were 40k, 3k and 3k samples of multi-turn dialogues, respectively. We adopted Exact Match (EM) and F1 score in SQuAD as metrics (Rajpurkar et al., 2016). Results in Table 1 show that while the utterance-based HA network is on par with established baselines, the proposed turn-based HA model obtains more gains, achieving the best EM and F1 scores.

Model	EM Score	F1 Score
Pointer LSTM	77.85	82.73
Bi-DAF	87.24	88.67
R-Net	88.93	90.41
Utterance-based HA	88.59	90.12
Turn-based HA (Proposed)	**91.07**	**92.39**

Table 1: Comparison with baseline models.

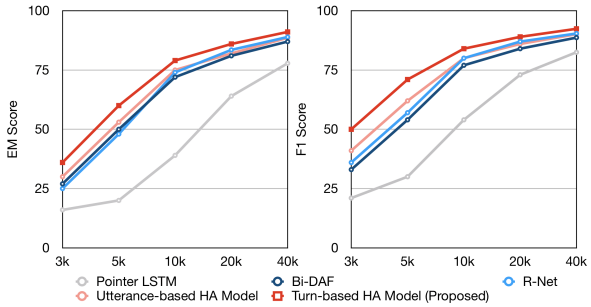

Figure 3: Results on different sizes of training data.

3.5 Evaluation in Low-Resource Scenarios

Limited amount of training data is a major pain point for dialogue-based tasks, as it is time-consuming and labor-intensive to collect and annotate natural dialogues at a large-scale. We expect the hierarchical structure to result in more efficient learning capabilities. We conducted experiments on a range of training sizes (from 3k to 40k) with a fixed-size test set (3k samples). As shown in Figure 3, the turn-based HA model outperforms all other models significantly when the training set is smaller than 20k.

3.6 Lengthy Sample Evaluation

Spoken conversations are often verbose with low information density scattered with topics not central to the main dialogue theme, especially since speakers chit-chat and get distracted during task-oriented discussions. To evaluate such scenarios, we adopted model-independent ADDSENT (Jia and Liang, 2017), where we randomly extracted sentences from SQuAD and inserted them before or after topically coherent segments. The average length of the augmented test set (3k samples), increased from 255 to 900. As shown in Table 2, the proposed turn-based model compares favorably when modeling lengthy dialogues.

Model	EM Score	F1 Score
Pointer LSTM	67.11 (-10.74)	72.67 (-10.06)
Bi-DAF	77.45 (-9.79)	79.55 (-9.12)
R-Net	79.96 (-8.97)	82.26 (-8.15)
Utterance-based HA	78.92 (-9.67)	80.72 (-9.40)
Turn-based HA	**85.25** (-5.82)	**87.18** (-5.21)

Table 2: Lengthy sample evaluation. Bracketed values denote absolute decrease of model performance in Section 3.6.

Model	EM Score	F1 Score
Pointer LSTM	60.99 (-16.86)	68.94 (-13.79)
Bi-DAF	74.58 (-12.66)	76.42 (-12.25)
R-Net	78.73 (-10.20)	80.38 (-10.03)
Utterance-based HA	77.84 (-10.75)	79.77 (-10.35)
Turn-based HA	**82.50** (-8.57)	**84.08** (-8.31)

Table 3: Out-of-distribution evaluation. Bracketed values denote absolute decrease of model performance in Section 3.7.

3.7 Out-of-Distribution Evaluation

Another evaluation was performed on an augmented set of dialogue samples, by adding three out-of-distribution symptom entities (bleeding, cold/flu, and sweating) to the corresponding conversations (3k samples). This was conducted on the well-trained models in Section 3.4. As shown in Table 3, the proposed turn-based HA model is the most robust in answering questions related to unseen symptoms/topics while till performing well on in-domain symptoms, thus showing potential generalization capabilities.

In summary, our overall experimental results demonstrate that the proposed hierarchical method achieves higher learning efficiency with robust performance. Moreover, the turn-based model significantly outperforms the utterance-based one, empirically verifying that it is appropriate to use turns as the basic semantic unit in coarse-grained attention for modeling dialogues.

4 Related Work

Machine comprehension of passages has achieved rapid progress lately, benefiting from large-scale datasets (Rajpurkar et al., 2016; Kocisky et al., 2018), semantic vector representations (Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019), and end-to-end neural modeling (Wang et al., 2017; Hu et al., 2018). The attention mechanism enables neural models to more flexibly focus on salient contextual segments (Luong et al., 2015; Vaswani et al., 2017), and is further improved by hierarchical designs for document processing tasks (Yang et al., 2016; Choi et al., 2017). Multi-level attention could be fused in hidden representations (Wang et al., 2017) or calculated explicitly (Hsu et al., 2018).

There is an established body of work studying how humans take turns speaking during conversations to better understand when and how to generate more natural dialogue responses (Sacks et al., 1974; Wilson et al., 1984; Schlangen, 2006). Utterance-level attention has also been applied to context modeling for different dialogue tasks such as dialogue generation (Serban et al., 2016) and state tracking (Dhingra et al., 2017). Recently, there is emerging interest in machine comprehension of dialogue content (Ma et al., 2018; Sun et al., 2019). To the best of our knowledge, our work is the first in exploiting turn-level attention in neural dialogue comprehension.

5 Conclusion

We proposed to comprehend dialogues by exploiting a hierarchical neural architecture through incorporating explicit turn-level attention scoring to complement word-level mechanisms. We conducted experiments on a corpus embodying verbal distractors inspired from real-world spoken dialogues that interrupt the coherent flow of conversation topics. Our model compares favorably to established baselines, performs better when there is limited training data, and is capable of addressing challenges from low information density of spoken dialogues and out-of-distribution samples.

Acknowledgements

This research was supported by funding for Digital Health and Deep Learning from the Institute for Infocomm Research (I2R) and the Science and Engineering Research Council (Project Nos. A1718g0045 and A1818g0044), A*STAR, Singapore. This work was conducted using resources and infrastructure provided by the Human Language Technology unit at I2R. We thank A. T. Aw, R. E. Banchs, L. F. D'Haro, P. Krishnaswamy, H. Lim, F. A. Suhaimi and S. Ramasamy at I2R, and W. L. Chow, A. Ng, H. C. Oh, S. Ong and S. C. Tong at Changi General Hospital for insightful discussions. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–220. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*. Association for Computational Linguistics.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–495, Vancouver, Canada. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141. Association for Computational Linguistics.

Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Ming Zhou. 2018. Read + verify: Machine reading comprehension with unanswerable questions. *CoRR*, abs/1808.05759.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Binte Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R. Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, Wai Leng Chow, and Nancy F. Chen. 2019. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2039–2048. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

David Schlangen. 2006. From reaction to prediction: Experiments with computational models of turn-taking. In *Ninth International Conference on Spoken Language Processing*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the 5th International Conference for Learning Representations*.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *CoRR*, abs/1902.00164.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *CoRR*, abs/1608.07905.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Thomas P Wilson, John M Wiemann, and Don H Zimmerman. 1984. Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, 3(3):159–183.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.