4.4 Tracking

The Tracking questions require more scattered information. A respondent should collect and accumulate specific information from a part of the passage, as the question is based on events happening repeatedly in a quarter or half of the game. For example, some questions ask about how many free-throws a player *A* will make in a quarter. As this figure does not appear in the passage, a respondent needs to count how many times the event '*A* makes a free-throw' occurs. In other words, it is necessary to track events relevant to the player '*A*' and 'free-throw'. When the player(*A*) is replaced with one team name, the new question is even more difficult because the information about each player belonging to the team should be tracked. Therefore, information tracking leads this kind of questions to be the most challenging ones in the dataset.

5 Baseline Models and Results

5.1 Models

To evaluate the QA performance on the LiveQA dataset, we implement 3 baseline models. The first is based on random selection, where the system randomly chooses a choice as the answer. The second is to choose the dominant option of each question. More concretely, 80.0% of questions are in format of 'yes' and 'no', where 57.8% has the answer 'no'. For the other multiple choice questions, 50.6% of them take the second option as the right answer. Thus, for 'yes/no' questions, we choose 'no', otherwise we choose the second option.

We also build a neural-network style baseline for our dataset to evaluate how state-of-the-art QA systems perform on the LiveQA dataset. Due to the uniqueness of our dataset, most of existing machine comprehension models are not suitable to it. For example, the QANet (Yu et al., 2018) model, which used to be a state-of-art model of SQuAD (Rajpurkar et al., 2016), is unavailable because it predicts the probability distribution of an answer's starting position and ending position in the context. But in LiveQA, a number of right answers do not directly appear in the context (e.g. an answer in format of 'can' or 'cannot'). Up to now, none of machine reading comprehension models has been designed for a dataset with consideration of timeline and mathematical computations. That means that the existing ones will not be likely to perform well on our dataset. The closest work to ours is multi-hop question answering, and thus we use a novel model Gated-Attention Reader (Dhingra et al., 2016) to experiment on LiveQA.

Gated-Attention Reader (GA) is an attention mechanism which uses multiplicative interactions between the query embedding and intermediate states of a recurrent neural network reader. GA enables a model to scan one document and the questions iteratively for multiple passes, and thus the multi-hop structure can target on most relevant parts of the document. It used to be the state-of-art model of several datasets, such as CNN/Daily Mail dataset (Hermann et al., 2015) and CBT dataset (Hill et al., 2015b).

The full context, which is usually composed of more than 1,000 sentences on average, is too heavy for GA as input. To apply GA to our dataset, we propose a pipeline method to first extract a set of candidate evidence sentences from the full content, and then apply the GA model on this set of sentences to predict the final answer. We employ TF-IDF style matching score to extract 50 most relevant sentences as the supporting evidence. To improve the accuracy of selecting the evidence candidates, if the question clearly requires some information after the game ends, we use the ending part of the content as the input.

Specifically, taken the embedding representation of a token, the Bi-directional Gated Recurrent Units (BiGRU) process the sequence in both forward and backward directions to produce two sequences of token-level representations, which are concatenated at the output as the final representation of the token. To perform multi-hop inference, the GA model reads the document and the query over k horizontal layers, where layer k receives the contextual embeddings $X_{(k-1)}$ of the document from the previous layer. At each layer, the document representation $D^{(k)}$ is computed by taking the full output of a document BiGRU where the previous layer embedding $X_{(k-1)}$ is the input. At the same time, a layer-specific query representation $Q^{(k)}$ is computed as the full output of a separate query BiGRU taking the query embedding Y as the input. The Gated-Attention is applied to $D^{(k)}$ and $Q^{(k)}$ to compute the contextual

embedding $X^{(k)}$.

$$X^{(k)} = GAttn(BiGRU(X^{(k-1)}), BiGRU(Y)) \tag{1}$$

After obtaining the query-awared document representation, we perform answer prediction by matching the similarity of answer and content. We use bidirectional Gated Recurrent Units to encode the candidate answers into vectors $A^(i)$, and then we compute matching score between summarized document and candidates using a bilinear attention. Finally we calculate the probability distribution of the options with softmax. The operations are similar to those in RACE (Lai et al., 2017).

$$s = softmax([Blin(A^i, D^{(k)});]_n^{i=1}) \tag{2}$$

5.2 Model Evaluation

Model	Acc
Random	50.0%
Dominant	**56.4%**
GA	53.1%

Table 4: The results of different baseline models on the test set. Random denotes randomly selecting an answer. Dominate denotes selecting the dominate option. GA denotes the gated-attention reader.

For the three baseline models, performance is reported with the accuracy on the test set in Table 4. The random selection method (Random) scores 50.0%, while the dominant option method (Dominate) reaches a score of 56.4%, which shows that our dataset does not have a certain pattern for the answers. Meanwhile, GA, which is a strong baseline for previous question answering problems, failed to perform better than the dominant option method and only achieves a score of 53.1%. Such results show that our dataset is challenging and needs further investigation for model design. In future work, how to incorporate temporal information and mathematical calculation into a QA model is the focus.

5.3 Case Study

In this subsection, we further analyze the prediction ability of the GA model. Table 5 shows some prediction cases in experimental results. From the first two questions, we can see that the model gives the correct answers when judging the result of a specific event. But for the other three questions which involve multiple events, the model fails to answer them correctly. A possible explanation is that, although GA is designed for multi-hop inference, it lacks ability in both information tracking and math calculation, which makes it difficult for the model to track down some complicated events.

We can see, for reading comprehension models that extract answers based on the similarity between the answer and the content, they would fail on LiveQA due to the fact that they cannot track down temporal information nor perform mathematical calculation. To outperform existing models on LiveQA, the system should consider focusing on tracking information of a certain event through the timeline. It should also have the ability to perform mathematical inference between different contents.

6 Conclusion

In this paper, we present LiveQA, a question answering dataset constructed from play-by-play live broadcast. LiveQA can evaluate a machine reading comprehension model in its ability to understand the timeline, track events and do mathematical calculation. It consists of 117k questions, which are time-dependent and need math inference. Due to the novel characteristics, it is hard for existing QA models to perform well on LiveQA. We expect our dataset will stimulate the development of more advanced machine comprehension models.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1057-1067, Hainan, China, October 31 - Novermber 1, 2020.

Question	Translation	Correct answer	Answer given by the model
跳球之争！本场比赛哪支球队获得第一轮进攻球权?	Jump ball fight! Which team will win the chance of the first round of offence?	勇士(The Warriors)	勇士(The Warriors)
湖人全场总得分是奇数还是偶数?	Will the total score of the Lakers at the end of the game be odd or even?	奇数(odd)	奇数(odd)
尼克杨第二节能否命中3分球?	Can Nick Young make a three pointer in the second quarter?	能(Yes)	不能(No)
第三节结束，76人能否领先湖人4分或更多?	At the end of the third quarter, Will the 76ers lead the Lakers by 4 points of more?	不能(No)	能(Yes)
谁先获得30分?	Who will score his 30th point earlier?	24分的哈登(James Harden who has got 24 points)	25分的托马斯(Isaiah Thomas who has got 25 points)

Table 5: Cases in the experimental results

Acknowledgement

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Natural Science Foundation Project of China (61876009), National Key Research and Development Project (2019YFB1704002), and National Social Science Foundation Project of China (18ZDA295). The corresponding author of this paper is Sujian Li.

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Fei Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision & Pattern Recognition*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015a. The goldilocks principle: Reading children's books with explicit memory representations. *Computer Science*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015b. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Tri Nguyen, Mir Rosenberg, Song Xia, Jianfeng Gao, and Deng Li. 2016. Ms marco: A human generated machine reading comprehension dataset.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Adam Trischler, Wang Tong, Xingdi Yuan, Justin Harris, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset.

Xiaojun Wan, Jianmin Zhang, Jin-ge Yao, and Tianming Wang. 2016. Overview of the nlpcc-iccpol 2016 shared task: Sports news generation from live webcast scripts. In Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng, editors, *Natural Language Understanding and Intelligent Applications*, pages 870–875, Cham. Springer International Publishing.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2017. Large-scale cloze test dataset created by teachers. *arXiv preprint arXiv:1711.03225*.

Jin-ge Yao, Jianmin Zhang, Xiaojun Wan, and Jianguo Xiao. 2017. Content selection for real-time sports news construction from commentary texts. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 31–40.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. 2016. Towards constructing sports news from live text commentary. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1361–1371.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1057-1067, Hainan, China, October 31 - Novermber 1, 2020.