A Multi-Task Learning Machine Reading Comprehension Model for Noisy Document (Student Abstract)

Zhijing Wu, Hua Xu*
[1]State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology(BNRist), Beijing 100084, China
wuzj18@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn

Abstract

Current neural models for Machine Reading Comprehension (MRC) have achieved successful performance in recent years. However, the model is too fragile and lack robustness to tackle the imperceptible adversarial perturbations to the input. In this work, we propose a multi-task learning MRC model with a hierarchical knowledge enrichment to further improve the robustness for noisy document. Our model follows a typical encode-align-decode framework. Additionally, we apply a hierarchical method of adding background knowledge into the model from coarse-to-fine to enhance the language representations. Besides, we optimize our model by jointly training the answer span and unanswerability prediction, aiming to improve the robustness to noise. Experiment results on benchmark datasets confirm the superiority of our method, and our method can achieve competitive performance compared with other strong baselines.

Introduction

Machine Reading Comprehension (MRC), making a machine to read a passage and answer a related question, is a challenging natural language processing task. There are a lot of successful end-to-end MRC models (Wang, Wu, and Yan 2018) that usually treat this task as a span prediction task and have an encoder-align-decoder neural structure. As for improving the robustness of the model, Hu et al. (2019) add specific answer verify module to determine whether the question is unanswerable or not. Wang and Jiang (2019) integrates the general background knowledge into the model.

However, there are still many shortcuts. Firstly, those models are very fragile since the performance will drop sharply when facing adversarially generated sentences without misleading humans. Besides, most models choose the segment with the highest probability as the answer, regardless of whether the question is answerable or not.

In this paper, we propose a multi-task learning MRC model with a hierarchical knowledge enrichment for noisy document. The experimental results on AddSent and AddOneSent datasets show that our method can achieve competitive performance than other strong baselines.

*Hua Xu is the corresponding author.

Method

Formally, given a query $Q = \{q_i\}_{i=1}^{m}$, a related passage $P = \{p_j\}_{j=1}^{n}$, q_i and p_j are original words, the task is to predict an answer $A = \{P_{ij}\}_{j=s}^{j=e}, 1 \leq s \leq e \leq n$ if the question is answerable. We propose a multi-task learning MRC model with knowledge enrichment block, and optimize it by jointly learning span prediction and unanswerable determination.

Contextual Encode Layer We generate passage and query contextual representations via a shared contextual encoder. Like many previous works, we gain the context embeddings at both word-level and char-level. For word-level, we use a pre-trained GloVe word embeddings and ELMo embeddings. We take the final hidden state of bi-RNN applied to learnable character vectors as char-level embeddings. Finally, we use a shared bi-LSTM on concatenated these three embeddings to get passage representation $A \in R^{n\times 2d}$ and query representation $B \in R^{m\times 2d}$.

Knowledge Enhancement Layer This layer can be thought of as a knowledge enhancement layer at different context levels, which utilizes coarse-to-fine background knowledge. It enhances the representation of a single sequence/layer or two sequences/layers.

We propose a hierarchical method of adding external knowledge into the contextual representation, based on a three-level addition strategy. At the low knowledge level, we follow (Wang and Jiang 2019) to determine the semantical connection between any two different words using a lexical English database WordNet. At the medium level, we choose the most useful knowledge sources and auxiliary skills as the most supportive knowledge based on the current query task to enhance the representation. At the high level, we apply a fusing strategy to enhance the document and query representation with the above helpful knowledge. After our enrichment operation, we get knowledgeable contextual passage representation $\hat{A} \in \mathbf{R}^{n\times 2d}$ and contextual query representation $\hat{B} \in \mathbf{R}^{m\times 2d}$.

Attention and Self-Attention Layer In order to find the answer in a passage, an important step is to match the passage and query, which aims to find the most relevant part in the passage. This attention layer employs an attention mech-

anism on query and passage representations to obtain co-dependent representations. The matching score between the i-th passage word and j-th query word is defined as follows:

$$S_{ij} = W_3[\hat{A}_i; \hat{B}_j; \hat{A}_i \odot \hat{B}_j; \hat{A}_i - \hat{B}_j] \in \mathbf{R}. \quad (1)$$

Then we apply knowledge enhancement strategy again to add the semantic and background knoeledge into co-dependent context representation. Finally, we use a self-attention layer to rewrite context representation, getting query-aware context representation $E \in R^{n \times 2d}$ that contains passage, query, and their correlation.

Answer Prediction Layer There are two sub-tasks in this prediction layer which share the above neural structure.
(1) Answer Span Prediction We employ pointer network from R-net (Wang et al. 2017) to calculate the probability distribution of the start and end index $p^{(1)}, p^{(2)} \in \mathbf{R}^n$:

$$s_j^t = \tanh(E_j W_h^P + h_{t-1} W_h^a) V_2 \quad (2)$$

$$p_i^{(t)} = \frac{exp(s_i^t)}{\sum_{i=1}^{n} exp(s_i^t)} \quad (3)$$

where h_0 is initialized with an attention-pooling vector of query representation.
(2) Unanswerability Prediction We add use sentinel identifier to determine the probability of unanswerability of the question $p^{(3)} \in \mathbf{R}^2$:

$$p^{(3)} = MeanPooling(E) V_3 \quad (4)$$

If $p_1^{(3)}$ is larger than the threshhold of unanswerability γ, the corresponding question is unanswerable.

Multi-Task Learning For training the model, we minimize the sum of negative log probabilities of the start and end indices and the unanswerability:

$$loss = -\sum_{j=1}^{N} (\log p_{y_j^s}^{(1)} + \log p_{y_j^e}^{(2)}) - \lambda \sum_{j=1}^{N} \log p^{(3)} \quad (5)$$

Experiments

Settings We combine the original SQuAD1.1 training set and the unanswerable samples in SQuAD2.0 about the same passage as our final training set. This operation ensures that we do not introduce additional query-passage-answer data.

Results We evaluate our method on these popular datasets, SQuAD1.1 and its two adversarial sets AddOne, AddOneSent. We compare our method with several strong baselines including FusionNet, SAN, R.M-Reader(Hu et al. 2018), QANet, SLQA(Wang, Wu, and Yan 2018), KAR(Wang and Jiang 2019). The performances of other baselines are taken directly from theri paper.

As the results are shown in Table 1, our model outperforms the state-of-the-art MRC models by a noticeable margin on two adversarial sets. It indicates that our model not only has a basic understanding and inferencing ability but also has a stronger anti-interference ability for noisy document. We also conduct an ablation study to evaluate the contribution of each model component, and the result shows that both knowledge enrichment and jointly training contribute to the model performance.

Single Model	Dev1.1 Set		AddSent	AddOneSent
	EM	F1	F1	F1
FusionNet	75.3	83.6	51.4	60.7
SAN	76.2	84.1	46.6	56.5
R.M-Reader	78.9	86.3	58.5	67.0
QANet	75.1	83.8	45.2	55.7
SLQA	**80.0**	**87.0**	54.8	64.2
KAR	76.7	84.9	60.1	72.3
Ours	73.9	82.4	**65.5**	**72.5**

Table 1: Results on SQuAD1.1 dev set and two of its adversarial dataset. Under the distraction of adversarially created sentences appended to the passage, our model still performs well and better than other competing baseline methods. Though we get lower result on dev set, it still performs better than several strong baselines like BiDAF (EM/F1 of 67.7%/77.3%).

Conclusion

In this paper, we propose an encode-align-decode framework of MRC model to tackle noisy documents. Firstly, we conduct a hierarchical knowledge enrichment to enhance the contextual representations by leveraging external knowledge from KBs. Secondly, we jointly train span prediction and unanswerability prediction tasks to improve the robustness of our MRC model. The experimental results show that our proposed model achieves significant performance improvement and become much more robust.

Acknowledgments This paper is funded by National Natural Science Foundation of China (Grant No: 61673235) and National Key R&D Program Projects of China (Grant No: 2018YFC1707605).

References

Hu, M.; Peng, Y.; Huang, Z.; Qiu, X.; Wei, F.; and Zhou, M. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, 4099–4106.

Hu, M.; Wei, F.; Peng, Y.; Huang, Z.; Yang, N.; and Li, D. 2019. Read + verify: Machine reading comprehension with unanswerable questions. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, 6529–6537.

Wang, C., and Jiang, H. 2019. Explicit utilization of general knowledge in machine reading comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 2263–2272.

Wang, W.; Yang, N.; Wei, F.; Chang, B.; and Zhou, M. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 189–198.

Wang, W.; Wu, C.; and Yan, M. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 1705–1714.