A Short Survey on Machine Reading Comprehension

Junyu Yin
s210027@e.ntu.edu.sg
Nanyang Technological University
Singapore

ABSTRACT

Machine reading comprehension (MRC) is a fundamental task in Natural Language Processing (NLP) for testing the capability of natural language understanding (NLU). It seeks to construct an intelligence agent which is capable of understanding text as people. With the recent booming of self-training methods in NLP, the emerging large-scale pre-trained language models (PLM) have caused a stir in the MRC community. However, as widely recognized, despite their powerfulness, there is still a long way to achieve the human-like reading ability. So in this literature review, I explore several research papers focusing on different aspects of the field to sketch this promising research direction.

KEYWORDS

machine reading comprehension, pre-trained language model, question answering

1 INTRODUCTION

MRC is a long-standing goal of NLU that aims to teach a machine to read and comprehend textual data [1]. As one of the major and challenging problems of NLP, it is very useful for multiple downstream tasks such as open domain question answering [18] and information retrieval [15].

Just like the language tests of humans, the most common way to test whether an intelligent agent (a person or an AI system) can fully understand a piece of text is to require her/him/it to answer related questions according to the given text. This task is formally defined as Question Answering (QA). In a typical task setting [1], an intelligent system is given a passage (context), a question, and asked to select a most appropriate answer from a list of candidate answers or to generate an answer directly.

Early MRC task was simplified as requiring systems to return a sentence that contains the right answer. The systems are based on rule-based heuristic methods, such as bag-of-words approaches [5], or manually generated rules [14]. With the recent emergence of large-scale pre-trained models [3, 7, 10, 19], a new paradigm of contextualized language representations was introduced to this area, which greatly strengthened the capacity of language encoder. As a result, the benchmark results of MRC were boosted remarkably.

Despite the powerfulness of these models, they can still fall behind the real human performance especially when facing some complicated tasks aiming to test the genuine reading ability. As a consequence, this field still needs to be develed into to bridge the performance gap between machine readers and real humans.

In this literature review, I will examine three papers with different focuses. One [17] investigates the potential of leveraging external knowledge to improve BERT [3] for MRC. Other two lay emphasis on complex logical reasoning over text [8] and how to select text sources to build new challenging benchmark dataset for MRC [16] respectively.

2 MOTIVATION

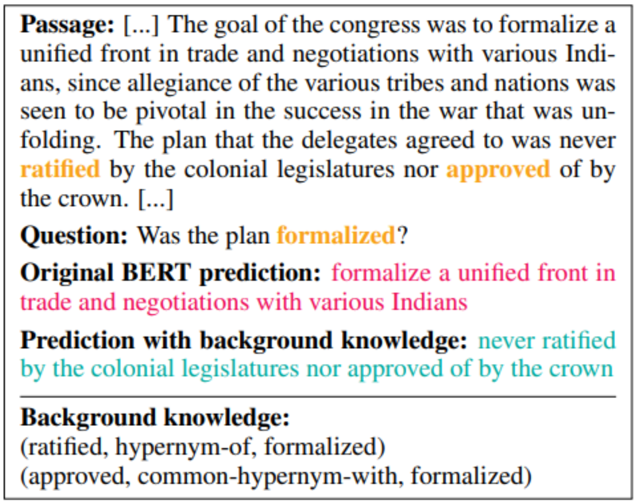
In the discussion that follows, I will use abbreviations for these three papers for brevity, that is, KT-NET for [17], AdaLoGN for [8] and WMRCQD for [16].

2.1 KT-NET

Thanks to the huge amounts of unlabeled corpus available and the sufficiently deep architectures used during pre-training, recent advanced PLMs are able to capture more complex linguistic phenomena and understand natural language better than before.

However, it is well known that true reading comprehension requires not only the understanding of the given passage but also some external knowledge to support text reasoning [2]. A motivating example is listed in Fig. 1 to show the importance and necessity of integrating background knowledge.

Thus, in [17], authors design KT-NET (short for Knowledge and Text fusion NET), a novel MRC method which improves PLMs with additional knowledge from knowledge bases (KBs). The aim is to take full advantage of both linguistic regularities covered by deep PLMs and high-quality knowledge derived from curated KBs to achieve better MRC.



Passage: [...] The goal of the congress was to formalize a unified front in trade and negotiations with various Indians, since allegiance of the various tribes and nations was seen to be pivotal in the success in the war that was unfolding. The plan that the delegates agreed to was never ratified by the colonial legislatures nor approved of by the crown. [...]

Question: Was the plan formalized?

Original BERT prediction: formalize a unified front in trade and negotiations with various Indians

Prediction with background knowledge: never ratified by the colonial legislatures nor approved of by the crown

Background knowledge:
(ratified, hypernym-of, formalized)
(approved, common-hypernym-with, formalized)

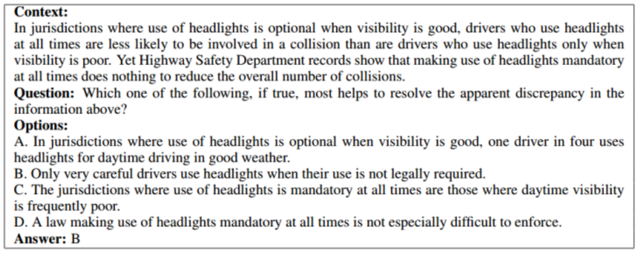
Figure 1: An example from SQuAD1.1 [13]. The vanilla BERT model fails to predict the correct answer. But it succeeds after integrating background knowledge collected from WordNet [12].

2.2 AdaLoGN

With the booming of PLMs, early MRC datasets are not difficult for state-of-the-art neural methods. And recent datasets [9, 20] are much more challenging for requiring understanding and reasoning over logical relations described in text, where neural models showed unsatisfactory performance.

Consider a motivating example shown in Fig. 2, with the help of propositional calculus, our humans can formalize propositions and then apply inference rules in propositional logic to get the correct answer. But how can machine readers solve such a task?

To meet the challenge, in AdaLoGN [8] researchers proposed a neural-symbolic approach which leverveges both merits of symbolic reasoners and neural models. It incorporates an adaptive logic graph network which adaptively infers logical relations to extend the graph and, essentially, realizes mutual and iterative reinforcement between neural and symbolic reasoning.



Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highway Safety Department records show that making use of headlights mandatory at all times does nothing to reduce the overall number of collisions.
Question: Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?
Options:
A. In jurisdictions where use of headlights is optional when visibility is good, one driver in four uses headlights for daytime driving in good weather.
B. Only very careful drivers use headlights when their use is not legally required.
C. The jurisdictions where use of headlights is mandatory at all times are those where daytime visibility is frequently poor.
D. A law making use of headlights mandatory at all times is not especially difficult to enforce.
Answer: B

Figure 2: An example from ReClor [20]. High-level logical reasoning skills are required to solve this problem.

2.3 WMRCQD

Recently state-of-the-art AI systems have shown performance comparable with humans on many early MRC datasets, which suggests that these benchmarks will no longer be able to measure future progress. So for a NLU benchmark dataset to be useful in research, it has to consist of examples that are diverse and difficult enough to discriminate among current and near-future state-of-the-art systems. To fulfill this goal, we will need to find better ways of building difficult datasets.

However, we do not have clear information on what aspects of text sources affect the difficulty and diversity of examples. Out of this motivation, authors of this paper [16] crowdsource multiple-choice reading comprehension questions to analyze what attributes of passages contribute to the difficulty and question types of the collected examples.

3 METHODOLOGY

3.1 KT-NET

In this paper, authors consider the extractive MRC task, that is, the answer to a given question is constrained as a contiguous span in the given passage. To solve this problem, the authors propose KT-NET, whose key idea is to enhance BERT with curated knowledge of KBs, thereby combining the advantages of both. The detailed architecture of the model is depicted in Fig. 3, with four major components: BERT encoding, knowledge integration, self-matching, and final output.

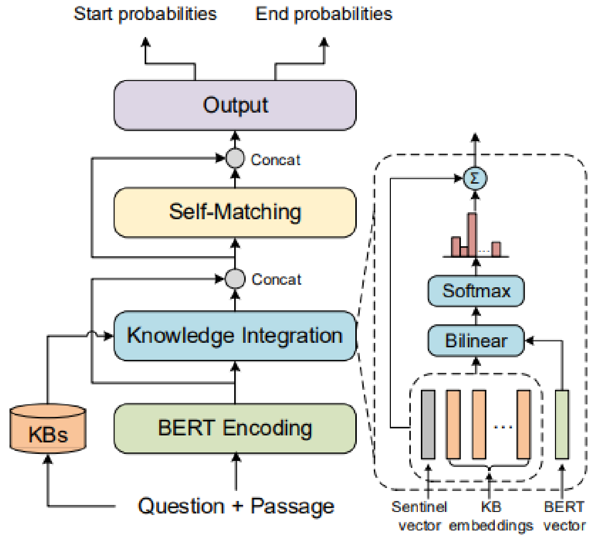


Figure 3: Overall architecture of KT-NET, with the knowledge integration module illustrated, cited from [17].

BERT Encoding Layer This layer uses BERT encoder to model passages and questions. Given a passage with m tokens $P = \{p_i\}_{i=1}^m$ and a question with n tokens $Q = \{q_j\}_{j=1}^n$, it computes for each token a context-aware representation. The final output of this layer is a set of hidden states $\{\mathbf{h}_i^L\}_{i=1}^{m+n+3}$, where L denotes the number of successive Transformer encoder blocks. The three new states are encoded from speicial tokens $\langle CLS \rangle$ and $\langle SEP \rangle$.

Knowledge Integration Layer This layer is designed to employ an attention mechanism to select the most relevant KB embeddings to further integrate knowledge into BERT, which makes the representations not only context-aware but also knowledge-aware.

Specifically, for each token s_i, we have its BERT representation $\mathbf{h}_i^L$ from the previous layer and retrieve a set of potentially relevant KB concepts $C(s_i)$. The retrieved KB embeddings $\{\mathbf{c}_j\}$ are aggregated by the attention mechanism into a single knowledge state vector $\mathbf{k}_i$ that encodes extra KB information w.r.t. the current token. This knowledge state vector $\mathbf{k}_i$ and BERT representation $\mathbf{h}_i^L$ are concatenated as output knowledge-enriched representations $\mathbf{u}_i$.

Self-Matching Layer This layer takes as input the knowledge-enriched representations $\{\mathbf{u}_i\}$, and employs a self-attention mechanism to further enable interactions among the context components $\{\mathbf{h}_i^L\}$ and knowledge components $\{\mathbf{k}_i\}$.

For a self-attention weight matrix A. we can compute for each token s_i an attended vector $\mathbf{v}_i = \sum_j a_{ij}\mathbf{u}_j$. Besides, authors also model the indirect interaction using a new self-attention weight matrix $\bar{A} = A^2$ and compute for each token s_i another attended vector $\bar{\mathbf{v}}_i = \sum_j \bar{a}_{ij}\mathbf{u}_j$. Finally, the output for each token is built by a concatenation $\mathbf{o}_i = [\mathbf{u}_i, \mathbf{v}_i, \mathbf{u}_i - \mathbf{v}_i, \mathbf{u}_i \odot \mathbf{v}_i, \bar{\mathbf{v}}_i, \mathbf{u}_i - \bar{\mathbf{v}}_i]$.

Output Layer This layer simply uses a linear output layer, followed by a standard softmax operation, to predict answer boundaries.

3.2 AdaLoGN

In this paper, authors focus on solving the multi-choice question answering task. Formally speaking, an input of this task is a 3-tuple $\langle c, q, O \rangle$ consists of a context c, a question q and a set of options O. Only one option in O is the correct answer to q given c. The goal of the task is to find this option.

The proposed solution to this task is outlined in Fig. 4. First, we enumerate each option $o \in O$ and generate the representations of c, q, o, that is, $\mathbf{g}_c, \mathbf{g}_q, \mathbf{g}_o$, the averaged vector representations for three token sequences, by RoBERTa. Second, we construct a raw text logic graph (TLG) where nodes $(u_1, \cdot, u_{|V|})$ represent EDUs [11] extracted from c, q, o and edges represent their logical relations. Then, with their initial representations obtained from the pre-trained language model, we adaptively extend the TLG (symbolic reasoning) and then pass messages (neural reasoning), in an iterative manner, to update node representations for generating the representation $\mathbf{h}_\mathrm{G}$ (by graph pooling) of the TLG. At last, we predict the correctness of o based on the above representations.

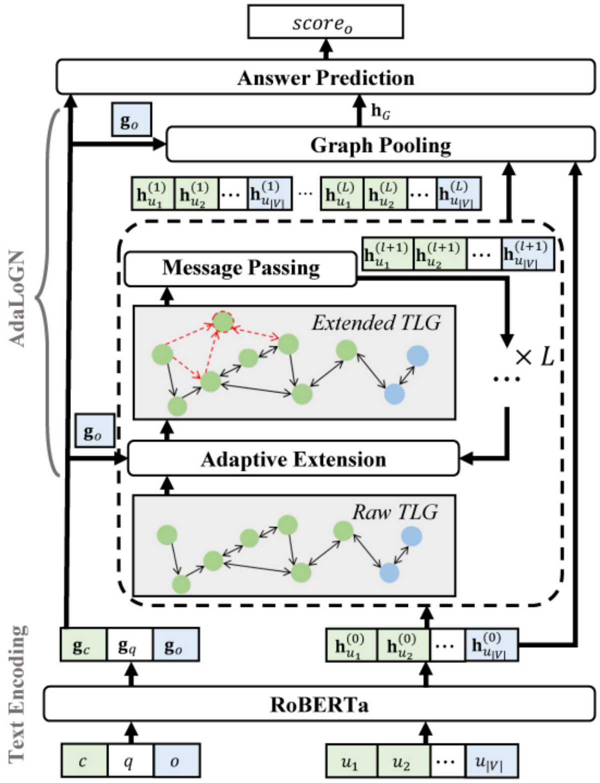

$score_o$
Answer Prediction
$\mathbf{g}_o$
$\mathbf{h}_G$
Graph Pooling
$\mathbf{h}_{u_1}^{(1)}$ $\mathbf{h}_{u_2}^{(1)}$ ··· $\mathbf{h}_{u_{|V|}}^{(1)}$ ··· $\mathbf{h}_{u_1}^{(L)}$ $\mathbf{h}_{u_2}^{(L)}$ ··· $\mathbf{h}_{u_{|V|}}^{(L)}$
Message Passing
$\mathbf{h}_{u_1}^{(l+1)}$ $\mathbf{h}_{u_2}^{(l+1)}$ ··· $\mathbf{h}_{u_{|V|}}^{(l+1)}$
Extended TLG
× L
AdaLoGN
$\mathbf{g}_o$
Adaptive Extension
Raw TLG
Text Encoding
$\mathbf{g}_c$ $\mathbf{g}_q$ $\mathbf{g}_o$
$\mathbf{h}_{u_1}^{(0)}$ $\mathbf{h}_{u_2}^{(0)}$ ··· $\mathbf{h}_{u_{|V|}}^{(0)}$
RoBERTa
c q o
u_1 u_2 ··· $u_{|V|}$

Figure 4: Overall architecture of AdaLoGN, cited from [8].

3.3 WMRCQD

This paper is kind different from other traditional papers of AI, focusing not on proposing a novel neural model to solve a given problem. Instead, it is more like a data analysis which aims to shed a light on how to build high-quality datasets.

In this work, authors analyze what kinds of passages make crowdsourced reading comprehension questions difficult. To collect difficult and high-quality examples, authors require crowdworkers to take a demanding qualification test. Then the results of those who pass this exam are used to represent the human performance. For machine performance, authors compute the average accuracy of eight different models from the following two classes: RoBERTa large [10] (four models with different random seeds) and DeBERTa large and xlarge [4]. Authors compute the difference between human and machine accuracy, using it as a measure of the question difficulty, to investigate whether there is a correlation between the question difficulty and linguistic aspects of the passage, such as their source, length, and readability.

To our surprise, authors find that the difficulty of collected questions does not depend on the differences of passages in linguistic aspects such as passage source, passage length, Flesch–Kincaid grade level [6], syntactic and lexical surprisal, elapsed time for answering, and the average word frequency in a passage. But the authors observe that questions that require numerical reasoning and logical reasoning are relatively difficult. In addition, the authors find several trends between the passage sources and reasoning types. For example, logical reasoning is more often required in questions written for technical passages, whereas understanding of a given passage's gestalt and the author's attitude toward it are more frequently required for argumentative and subjective passages than expository passages.

4 CONCLUSION

In this assignment, I conduct a concise survey about machine reading comprehension. The three papers I've read address the issues of bringing in external knowledge, logical reasoning, and how to structure datasets, respectively. Furthermore, these studies reveal that there is a long way for state-of-the-art machine readers to achieve human-like performance. As a promising research direction, we still have a lot things to do. A direct idea can be combining these papers together. For example, we first construct a dataset that requires external knowledge other than given articles and some reasoning skills to answer questions. Then we combine the above-mentioned models, that is, KBs, symbolic logic reasoner and neural model together to solve such problem.

REFERENCES

[1] Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2020. A survey on machine reading comprehension systems. *Natural Language Engineering* (2020), 1–50.

[2] Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2358–2367.

[3] 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* (2018).

[4] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

[5] Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. 325–332.

[6] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Technical Report. Naval Technical Training Command Millington TN Research Branch.

[7] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[8] Xiao Li, Gong Cheng, Ziheng Chen, Yawei Sun, and Yuzhong Qu. 2022. AdaLoGN: Adaptive Logic Graph Network for Reasoning-Based Machine Reading Comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7147–7161.

[9] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124* (2020).

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[11] William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse* 8, 3 (1988), 243–281.

[12] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[13] 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*. 2383–2392.

[14] Ellen Riloff and Michael Thelen. 2000. A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*.

[15] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.

[16] Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. 2022. What Makes Reading Comprehension Questions Difficult?. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6951–6971.

[17] An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2346–2357.

[18] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2013–2018.

[19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[20] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2019. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations*.