**VICTORIA UNIVERSITY**

**Encoding knowledge in SQuAD 2.0 Dataset**

**NEF6101 – Research Thesis**

**Proposal & Literature Review**

**Sigdel, Bijay**

**4603843**

Table of Contents

[Literature Review 2](#_Toc17593594)

[Related Work (Focused Reading and Reviews) 4](#_Toc17593595)

[Research Problem 8](#_Toc17593596)

[Research Method 9](#_Toc17593597)

[References 12](#_Toc17593598)

[Appendix 14](#_Toc17593599)

# 

# **Literature Review**

**Encoding Knowledge in SQuAD 2.0 Dataset**

This research is proposed to enhance SQuAD 2.0 dataset. SQuAD 2.0 (Stanford Question Answering Dataset) is a dataset consisting of articles and questions and answers based on these articles. SQuAD 2.0 was created out of crowdsourcing, where crowd workers were asked to create unanswerable questions from articles in SQuAD dataset (older version), while referencing entities in the paragraph and ensuring that a plausible answer is present. This dataset can be used to train learning algorithms like BERT (Bidirectional Encoder Representations from Transformers) and fine-tuning them. It has been extensively used in the research community for evaluation purposes. Searching with keyword “SQuAD 2.0” on IEEE Xplore returns one result which reveals that research on SQuAD 2.0 dataset is still very new.

Searching keyword “SQuAD 2.0” on Google Scholar returns 1680 results. Following are the list of articles related to SQuAD 2.0 dataset:

1. Know What You Don't Know: Unanswerable Questions for SQuAD
2. Stochastic answer networks for squad 2.0
3. Exploring Neural Net Augmentation to BERT for Question Answering on SQUAD 2.0
4. A qualitative comparison of coqa, squad 2.0 and quac
5. pair2vec: Compositional word-pair embeddings for cross-sentence inference
6. A bert baseline for the natural questions
7. Read+ verify: Machine reading comprehension with unanswerable questions
8. Sogou Machine Reading Comprehension Toolkit
9. Learning to Ask Unanswerable Questions for Machine Reading Comprehension
10. A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension
11. Unified Language Model Pre-training for Natural Language Understanding and Generation
12. QADiver: Interactive Framework for Diagnosing QA Models
13. CFO: A Framework for Building Production NLP Systems
14. Modified VS3-NET for Reading Comprehension and Question Answering with No-Answers
15. U-Net: Machine Reading Comprehension with Unanswerable Questions
16. Contextual Aware Joint Probability Model Towards Question Answering System
17. Bert: Pre-training of deep bidirectional transformers for language understanding
18. EQuANt (Enhanced Question Answer Network)
19. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions
20. Quac: Question answering in context

It shows research articles on question answering dataset, implementing BERT on SQuAD 2.0 dataset, augmenting BERT, and others. “Know What You Don't Know: Unanswerable Questions for SQuAD” is award winning article which describes in depth about SQuAD 2.0 dataset. Not only does this article show fundamentals of SQuAD 2.0 dataset, but it also helps to identify SQuAD 2.0 dataset’s weaknesses. So, this article is chosen for focused reading.

To refine the research, keyword “enhancing SQuAD 2.0 dataset” was used in google scholar. Following results were returned:

1. Know What You Don't Know: Unanswerable Questions for SQuAD
2. Read+ verify: Machine reading comprehension with unanswerable questions
3. Stochastic answer networks for squad 2.0
4. EQuANt (Enhanced Question Answer Network)
5. Exploring Neural Net Augmentation to BERT for Question Answering on SQUAD 2.0
6. BLCU\_NLP at SemEval-2019 Task 8: A Contextual Knowledge-enhanced GPT Model for Fact Checking
7. SG-Net: Syntax-Guided Machine Reading Comprehension
8. Learning to Ask Unanswerable Questions for Machine Reading Comprehension
9. Encoding Knowledge Graph with Graph CNN for Question Answering
10. Contextual Aware Joint Probability Model Towards Question Answering System

Article “Encoding Knowledge Graph with Graph CNN for Question Answering” propose encoding external knowledge to question answering system. This article is also chosen for focused reading.

BERT is a reading comprehension model which applies bidirectional training to have deeper sense of language. It looks at text sequence from both left and right direction. It learns the context of word based on all of its surrounding. To achieve this, BERT employs two strategies. They are:

1. Masked LM: Before feeding word sequences into BERT, 15% of the words in each sequence are replace by mask. The model then attempts to predict the original value of the masked words, based on the other non-masked words in the sequence.
2. Next Sequence Prediction: During training, 50% of the inputs are a pair in which the second sentence is the subsequent sentence in the original document, while in the other 50% a random sentence from the corpus is chosen as the second sentence. BERT attempts to distinguish these sentences.

# **Related Work (Focused Reading and Reviews)**

1. **Know What You Don’t Know: Unanswerable Questions for SQuAD**

(Rajpurkar, Jia & Liang, 2018) presented SQuAD 2.0, the newer version of the Stanford Question Answering Dataset (SQuAD). SQuAD 2.0 combines SQuAD data with more than 50,000 unanswerable questions. They employed crowdworkers to create these unanswerable questions. The questions were built with two main requirements. They are:

* Unanswerable questions should be relevant to the paragraph’s context
* Paragraph’s context should present some span whose type resembles the question type

Authors also employed other crowdworkers to answer questions in SQuAD 2.0 dataset and the final answer was chosen by majority vote. They trained and tested this dataset on three different models and found out best model (DocQA + ELMo) achieves only 66.3 F1 score. However, SQuAD 1.1 obtains 85.8 F1 on the same model.

This proves a larger variation between SQuAD 2.0 and SQuAD 1.1 dataset making SQuAD 2.0 a more challenging, diverse and large-scale dataset.

Why did the authors conduct the research?

SQuAD, earlier version of SQuAD 2.0 doesn’t have any unanswerable questions. It focused only on questions that has guaranteed answer existing in the passage. So, models can do well on SQuAD by learning context and type-matching heuristics. Because of this, some research even produced systems that surpassed human-level accuracy. This doesn’t mean the model were able to understand language at deeper level.

Other datasets like Zero-shot Relation Extraction and NewsQA data collection process, although they have unanswerable questions, they were still lacking plausible answer or were out-of-scope. So, authors conducted the research.

What did the authors do? What innovation they have and what are the key outcomes?

SQuAD 2.0 consists of over 50,000 unanswerable questions. They hired crowdworkers to craft all these questions. For every passage in an article, they asked workers to create up to five questions that are impossible to answer. Each of these questions supports two main criteria i.e. plausibility and relevancy. Author also removed questions from workers who wrote 25 or fewer questions on an article. They also employed crowdworkers for answering all these questions and defining the human accuracy. Unlike SQuAD, SQuAD 2.0 gave more importance to human accuracy by selecting the final answer of each question by majority vote.

They trained and tested this dataset on three different models namely BiDAF-No-Answer (BNA), and two versions of the DocumentQA No-Answer (DocQA) model and found that the best model is DocQA + ELMo achieving only 66.3 F1 score which is 23.2 points lesser then human accuracy of 89.5 F1. While on the same model, SQuAD 1.1 obtains 85.8 F1, only 5.4 points worse than humans. This provided robust challenge to all reading comprehension models.

What is the problem in their research?

Authors present a comprehensive dataset that is difficult to answer. We, human beings are benefitted by general knowledge that we have collected since the very childhood. However, machines do not have general knowledge and its more challenging for them. The dataset is not challenging in a sense that it captures depth of the passage, but in a sense, it has tweaked some of the words in the question with their synonyms, literal meanings and requires some general knowledge to solve.

What can be further improved?

The questionnaires in SQuAD 2.0 can be fitted with some general knowledge. This will help the models to grasp the context of the passage. Along with providing general knowledge, questions with more in-depth context of passage context can be added.

1. **Encoding Knowledge Graph with Graph CNN for Question Answering**

(Laugier, Wang, Foo, Guenais & Chandrasekhar, 2019) proposed general framework which incorporates external knowledge (encoded as knowledge graphs) into question answering systems using graph Convolutional Neural Networks (CNN). They implemented this framework on Stochastic Answer Networks (SAN) producing an improved F1 score and EM score.

The authors build knowledge sub-graphs from ConceptNet for the passage context associated with each question. They developed knowledge-aware context representation in the base model i.e. Stochastic Answer Networks (SAN) and augmented the question answering capability of the model.

They analyse the framework in different depths of Graph CNN. They also examine the framework in single-layer graph structure and different concept embedding initializations.

Why did the authors conduct the research?

Authors believe that because of the ambiguity of natural language and it’s implied context, question answering requires external knowledge to resolve. They claim that previous research has implemented knowledge graphs for question answering, however, they haven’t sufficiently used the graph structure of knowledge graph that is fundamental for modelling relationships

between the concepts.

What did the authors do? What innovation they have and what are the key outcomes?

They utilize Graph Convolutional Neural Networks (Graph CNNs) to represent knowledge from ConceptNet. They build knowledge sub-graph including relevant concepts and relations in order to model knowledge related to the context and produce a context-related knowledge vector. This knowledge vector is utilized in question answering model to obtain knowledge-aware context representation and hence, answer the question with good accuracy.

They test the effects of using different depths of Graph CNN. They also implemented the framework with a single-layer graph structure to show the importance of exploiting the graph structure. Moreover, they also implemented different concepts embedding initializations, which are random initialization, GloVe 100 GloVe 300 and NumberBatch confirming the importance of appropriate parameter initialization for vertices in Graph CNN.

This framework was successfully applied on Stochastic Answer Networks (SAN) and significant improvement was carried out. This can be shown in figure below:

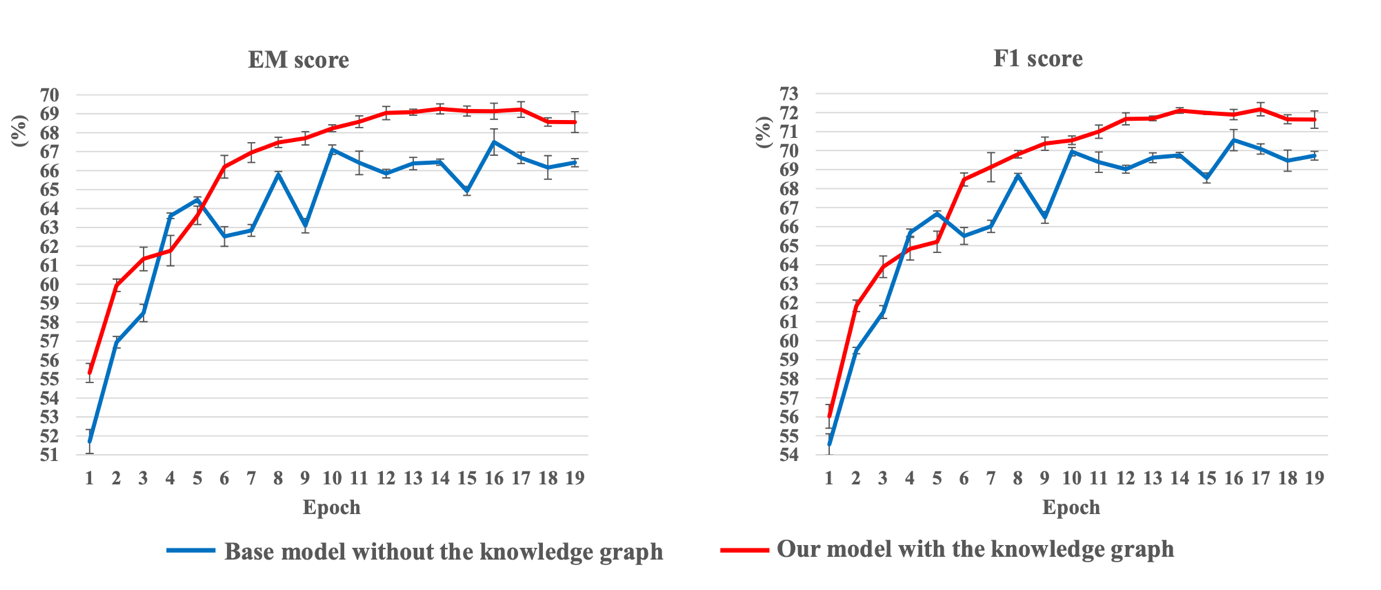


Figure 1: Picture Showing EM and F1 score in base model (SAN)

What is the problem in their research?

Although authors claim that they have developed a general knowledge-aware framework for question answering, there is no available result showing its implementation and improvement in other popular dataset like SQuAD 2.0. Moreover, their experiment shows that deeper graph CNN structure can encode knowledge of sub-graphs better, however, such implementation is hindered by GPU limitation. Therefore, this research is not feasible for general Natural Language Processing (NLP).

What can be further improved?

Rather than utilizing deeper graph CNN structure which doesn’t seem to be feasible for general purpose because of its limitations, authors could have implemented other approaches that are not bounded by processing and price limitations. Authors could implement this approach in different dataset before setting it as a general module.

# **Research Problem**

SQuAD 2.0 dataset has over 50,000 unanswerable questions. Researchers have employed synonyms, literal meanings, negations or some kinds of general knowledge to make these questions unanswerable. These questions do not provide general knowledge required for answering the question which is one problem earlier research has.

One research paper, “Encoding Knowledge Graph with Graph CNN for Question Answering” has proposed encoding external knowledge for question answering and has improved the F1 and EM score in Stochastic Answer Networks (SAN). However, there is no any experiment showing its implementation and enhancement in SQuAD 2.0. Moreover, for achieving better performance, this research has its own drawbacks like higher GPU requirements. This is another problem previous research has.

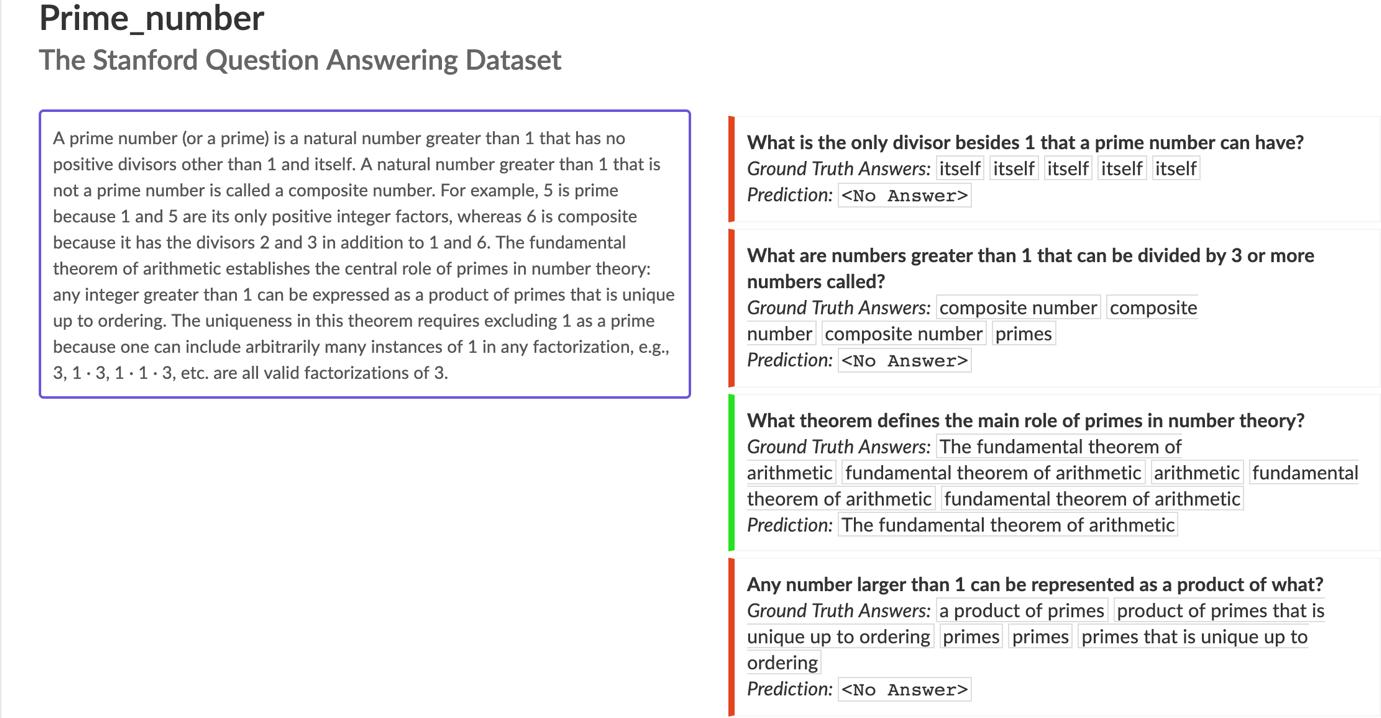
****

Figure 2: Picture taken from prime number article and its questionnaire in SQuAD 2.0 Dataset

In this picture, the first question in the right-hand side is extracted from the first line of the paragraph in the left-hand side. The question simply replaces besides word in the passage with other than. In English language, ‘besides’ has the same meaning as ‘other than’. Machine can’t figure this simple general knowledge and is unable to answer the question.

# **Research Method**

This research proposes to implement general knowledge in SQuAD 2.0 dataset. The general knowledge will be provided alongside question and it will be relevant to the context in the passage and the question. Machine will apply this general knowledge to answer the question.

One sample article and questionnaires based on the article are provided below. General knowledge is provided alongside each question.

**Australia Our Country** (Sample information/article taken fromhttps://www.australia.gov.au/about-australia/our-country)

Paragraph 1: Population

As of December 2014, Australia's population is roughly 23.6 million people. The most populous states are New South Wales and Victoria, with their respective capitals, Sydney and Melbourne, the largest cities in Australia.

Australia's population is concentrated along the coastal region of Australia from Adelaide to Cairns, with a small concentration around Perth, Western Australia. The centre of Australia is sparsely populated.

Questions

1. What is Australia’s population in late 2014?

Ans: 23.6 million

General Knowledge: late 2014 similar December 2014

1. Which are the most occupied states in Australia?

Ans: New South Wales and Victoria

General Knowledge: Populous synonym Occupied

1. Along which region is Australia highly populated?

Ans: Coastal region

General Knowledge: concentrated similar populated

Paragraph 2: Geography

Australia is an island continent and the world's sixth largest country (7,682,300 sq km).

Lying between the Indian and Pacific oceans, the country is approximately 4,000 km from east to west and 3,200 km from north to south, with a coastline 36,735 km long.

Canberra is Australia's capital city. With a population of approximately 380,000 people and situated in the Australian Capital Territory, Canberra is roughly halfway between the two largest cities Melbourne and Sydney.

Australia has 19 listed World Heritage properties. Australia is also famous for its landmark buildings including the Sydney Harbour Bridge; its ancient geology, as well as for its high country.

Questions

1. Which country is the world’s sixth largest country?

Ans: Australia

Note: It is a simple question

1. Australia is between which oceans?

Ans: Indian and Pacific oceans

General Knowledge: lying between similar is between

1. Is Canberra Australia’s capital city?

Ans: Yes

Note: It is a simple question

1. Is the population of Canberra 0.38 million?

Ans: Yes

Note: 380,000 equal to 0.38 million

Paragraph 3: History

Australia's first inhabitants, the Aboriginal people, are believed to have migrated from some unknown point in Asia to Australia between 50,000 and 60,000 years ago.

While Captain James Cook is credited with Australia's European discovery in 1770, a Portuguese possibly first sighted the country, while the Dutch are known to have explored the coastal regions in the 1640s.

The first European settlement of Australia was in January 1788, when the First Fleet sailed into Botany Bay under the command of Captain Arthur Phillip. Originally established as a penal colony, by the 1830s the number of free settlers was increasing. Transportation of convicts to the eastern colonies was abolished in 1852 and to the western colonies in 1868.

Questions

1. Who are the Australia’s first residents?

Ans: Aboriginal people

General Knowledge: residents synonym inhabitants

1. Who is honoured for Australia’s European discovery?

Ans: Captain James Cook

General Knowledge: honoured similar credited

1. Who are known to have explored the coastal regions in Australia in the mid 1600s?

Ans: Dutch

General Knowledge: mid 1600s equal to 1640/1650/1660

1. When was the first African settlement of Australia?

Ans: No answer

Note: African means people from africa

The general knowledge provided represents different type of knowledges. They are synonyms, similar to, equal to, antonyms, mathematical calculations, measurement unit conversion and others. These knowledges can be classified as per their types and usage.

General Knowledge functionality, on one hand will urge the researchers to create more in-depth questionnaire. On the other hand, it will challenge the reading comprehension models to get the in-depth knowledge of the language and answer the question.

# **References**

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Horev, R. (2018). BERT Explained: State of the art language model for NLP. [online] Medium. Available at: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 [Accessed 24 Aug. 2019].

Konrád, J. (2018). Transfer learning for question answering on SQuAD. Prague May, 10, 1.

Laugier, L., Wang, A., Foo, C.-S., Guenais, T., & Chandrasekhar, V. ENCODING KNOWLEDGE GRAPH WITH GRAPH CNN FOR QUESTION ANSWERING.

Liu, X., Li, W., Fang, Y., Kim, A., Duh, K., & Gao, J. (2018). Stochastic answer networks for squad 2.0. arXiv preprint arXiv:1809.09194.

narayanan, d. (2018). SQuADRUn — SQuAD 2.0 from Stanford. [online] Medium. Available at: https://medium.com/@deepaknarayanan120199/squadrun-squad-2-0-from-stanford-c620b12918a4 [Accessed 24 Aug. 2019].

Rajpurkar, P., Jia, R. and Liang, P. (2018). Know What You Don’t Know: Unanswerable Questions for SQuAD.

Yatskar, M. (2018). A qualitative comparison of coqa, squad 2.0 and quac. arXiv preprint arXiv:1809.10735.

# **Appendix**

**Appendix A.1 Literature search of SQuAD 2.0**

This research is proposed to enhance SQuAD 2.0 dataset. The initial keywords selected was “SQuAD 2.0 Dataset”

The research database searched was IEEE Explore. It returned only one result as follows:

1. Modified VS3-NET for Reading Comprehension and Question Answering with No-Answers

Then Google Scholar was used as the research database and it returned 1680 results. The title of first 20 articles are as follows:

1. Know What You Don't Know: Unanswerable Questions for SQuAD
2. Stochastic answer networks for squad 2.0
3. Exploring Neural Net Augmentation to BERT for Question Answering on SQUAD 2.0
4. A qualitative comparison of coqa, squad 2.0 and quac
5. pair2vec: Compositional word-pair embeddings for cross-sentence inference
6. A bert baseline for the natural questions
7. Read+ verify: Machine reading comprehension with unanswerable questions
8. Sogou Machine Reading Comprehension Toolkit
9. Learning to Ask Unanswerable Questions for Machine Reading Comprehension
10. A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension
11. Unified Language Model Pre-training for Natural Language Understanding and Generation
12. QADiver: Interactive Framework for Diagnosing QA Models
13. CFO: A Framework for Building Production NLP Systems
14. Modified VS3-NET for Reading Comprehension and Question Answering with No-Answers
15. U-Net: Machine Reading Comprehension with Unanswerable Questions
16. Contextual Aware Joint Probability Model Towards Question Answering System
17. Bert: Pre-training of deep bidirectional transformers for language understanding
18. EQuANt (Enhanced Question Answer Network)
19. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions
20. Quac: Question answering in context

Review -There are two observations by reviewing the outcomes:

1. The research on SQuAD 2.0 is new and active.
2. These research articles are focused on question answering dataset, implementing BERT on SQuAD 2.0 dataset, augmenting BERT, and others

**Appendix A.2 Literature search of Enhancing SQuAD 2.0 Dataset**

To focus the research on improving SQuAD 2.0 dataset, search keywords selected was “Enhancing SQuAD 2.0 Dataset”.

The returned results were 1130.

The titles of the first 20 articles are listed as follows:

1. Know What You Don't Know: Unanswerable Questions for SQuAD
2. Read+ verify: Machine reading comprehension with unanswerable questions
3. Stochastic answer networks for squad 2.0
4. EQuANt (Enhanced Question Answer Network)
5. Exploring Neural Net Augmentation to BERT for Question Answering on SQUAD 2.0
6. BLCU\_NLP at SemEval-2019 Task 8: A Contextual Knowledge-enhanced GPT Model for Fact Checking
7. SG-Net: Syntax-Guided Machine Reading Comprehension
8. Learning to Ask Unanswerable Questions for Machine Reading Comprehension
9. Encoding Knowledge Graph with Graph CNN for Question Answering
10. Contextual Aware Joint Probability Model Towards Question Answering System
11. Influence of Sentiment in Question Answering Systems
12. pair2vec: Compositional word-pair embeddings for cross-sentence inference
13. SpanBERT: Improving Pre-training by Representing and Predicting Spans
14. Sogou Machine Reading Comprehension Toolkit
15. Explicit Utilization of General Knowledge in Machine Reading Comprehension
16. Multi-task Learning with Sample Re-weighting for Machine Reading Comprehension
17. Using Natural Language Relations between Answer Choices for Machine Comprehension
18. ERNIE: Enhanced Language Representation with Informative Entities
19. Summarization Filter: Consider More About the Whole Query in Machine Comprehension
20. Short Text Conversation Based on Deep Neural Network and Analysis on Evaluation Measures

Review -There are two observations by reviewing the outcomes:

1. These research articles focus on improving question answering dataset.
2. “Encoding Knowledge Graph with Graph CNN for Question Answering” propose encoding external knowledge to question answering system. This article is chosen for focused reading.