FreebaseQA: A New Factoid QA Data Set Matching Trivia-Style Question-Answer Pairs with Freebase

Kelvin Jiang[*] [†] and **Dekun Wu**[*] and **Hui Jiang**
Department of Electrical Engineering and Computer Science
York University, 4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada
{kelvin,jackwu,hj}@eecs.yorku.ca

Abstract

In this paper, we present a new data set, named *FreebaseQA*, for open-domain factoid question answering (QA) tasks over structured knowledge bases, like Freebase. The data set is generated by matching trivia-type question-answer pairs with subject-predicate-object triples in Freebase. For each collected question-answer pair, we first tag all entities in each question and search for relevant predicates that bridge a tagged entity with the answer in Freebase. Finally, human annotation is used to remove false positives in these matched triples. Using this method, we are able to efficiently generate over 54K matches from about 28K unique questions with minimal cost. Our analysis shows that this data set is suitable for model training in factoid QA tasks since FreebaseQA provides more linguistically sophisticated questions than other existing data sets. The data set is available for free download at http://github.com/infinitecold/FreebaseQA.

1 Introduction

Within the field of natural language processing (NLP), there has been an increase in developments towards various real-world applications, such as factoid question answering (QA): the process of obtaining the answer(s) to a factual question - similar to trivia game settings. For this task to be successfully completed, there are several steps that need to occur. Notably, we need to interpret and parse the question, determine the domain of relevance, eliminate ambiguities if existent, and pinpoint the exact answer to the question asked. Fortunately, this task has been simplified with the emergence of large knowledge graphs, including Freebase (Bollacker et al., 2008), from where we can retrieve information. Knowledge graphs are colossal networks of data that describe concepts, entities, and their relations. In fact, Freebase is the largest publicly-available knowledge graph, consisting of 4 million nodes and approximately 3 billion edges (Google, 2017). Each node represents an entity existing in the physical world, such as a person, a location, or an organization. Each edge represents a relation between two entities, in a directed manner from a *subject* node to an *object* node. In Freebase, these edges are referred to as *predicates*, and a collection of a subject-predicate-object is referred to as a *triple*. An example triple in Freebase is the subject `Clarissa`, predicate `book.written_work.author` and object `Samuel Richardson`, explaining that the book Clarissa is written by author Samuel Richardson. Specifically, we take advantage of these relations between entities, which describe facts, to help with factoid question answering. We believe open-domain factoid QA over structured knowledge graphs like Freebase is a very interesting NLP task since it opens up many interesting real-world applications, such as natural language based query and search. Finally, once questions are formulated using a variety of rich and sophisticated representations in natural languages, such factoid QA tasks may serve as an excellent testbed to study many natural language understanding problems, e.g., examine the recently emerging research efforts to combine neural models with the traditional symbolic processing methods (Liang et al., 2017; Mou et al., 2017).

On the other hand, machine learning approaches for NLP are data hungry since they require large amounts of real-world data to train the models for the best possible performance. Existing data sets for the factoid QA task over structured knowledge bases are either too small in scale to train neural networks effectively, or contain questions

[*] Equal contribution.
[†] Currently at University of Waterloo. Work was done at York University.

Proceedings of NAACL-HLT 2019, pages 318–323
Minneapolis, Minnesota, June 2 - June 7, 2019.

that are too simple in linguistic structure to amply cover real-world scenarios. In this paper, we introduce a new data set for open-domain QA over Freebase, called *FreebaseQA*, which is created by matching trivia-type question-answer pairs with Freebase triples that reflect the semantic meaning of the questions. FreebaseQA contains over 54K matches from about 28K unique questions that can be used to train machine learning (ML) models and help the development of factoid QA systems for more realistic applications. Particularly, these matches may be used to train ML models to align natural language questions with Freebase predicates to search for the correct answers in Freebase. Our analysis shows that FreebaseQA provides an advantage over all pre-existing data sets with similar objectives, which are either too small or only contain questions that are too simple in linguistic structure. These results will be explained in detail in Section 4.

2 Related Work

Factoid QA data sets involving question-answer pairs as well as their corresponding Freebase matches have been created in the past. In (Berant et al., 2013), factoid QA over knowledge graphs are formulated as semantic parsing problems, where each natural language question is first converted into a logic form to retrieve the answer with traditional symbolic approaches. In (Berant et al., 2013), a small-scale data set of several thousands of question-answer pairs, called WebQuestions, is created by human annotators. In (Yih et al., 2016), the WebQuestions set is further refined by providing human-annotated semantic parses for some questions that are answerable using Freebase, which is called WebQuestionsSP (WebQSP). Recently, deep learning approaches have become popular in the field of NLP. Neural networks require far more training data than a small data set of several thousands of samples. In (Bordes et al., 2015), a much larger QA data set of about 100K question/answer pairs, called SimpleQuestions, is created. In this work, some randomly chosen Freebase triples are shown to human annotators. For each given triple, an annotator is asked to manually compose a question to reflect the relation in the triple. The issues with SimpleQuestions lie in that most constructed questions are quite simple in linguistic structure and many questions even directly use the keywords in the Freebase predicates since human annotators may be greatly limited in composition when a particular triple is shown. According to (Petrochuk and Zettlemoyer, 2018), SimpleQuestions is nearly solved with only standard neural network methods if its linguistic ambiguity is taken into account. In (Vlad Serban et al., 2016), a large QA data set is automatically generated by neural networks but it obviously lacks rich linguistic variations. Additionally, many similar factoid QA data sets are also released for other non-English languages, e.g. WebQA in (Li et al., 2016). Meanwhile, another direction of data collection efforts involve QA in various reading comprehension tasks, e.g. SQuAD in (Rajpurkar et al., 2016), MS-MARCO in (Nguyen et al., 2016), TriviaQA in (Joshi et al., 2017). However, we believe question answering over structured knowledge graphs remains a viable NLP task for the promising research direction to combine neural computing methods with the traditional symbolic processing approaches.

3 Constructing the FreebaseQA Data Set

In this section, we outline the construction procedure of the FreebaseQA data set, which consists of about 54K matches in the form of two examples shown in Table 1.

3.1 Preparation of Question-Answer Pairs

In FreebaseQA, we have not generated any new question-answer pairs but we have instead collected pre-composed trivia-type factoid questions from a number of sources. Unlike SimpleQuestions, these questions are independently composed for human contestants in various trivia-like competitions. As a result, these questions show much richer linguistic variation and complexity than almost all existing KB-QA data sets. In particular, we use the TriviaQA (Joshi et al., 2017) data set as the primary source of our QA pairs, while also including questions scraped from the trivia websites, KnowQuiz (`http://www.knowquiz.com`), QuizBalls (`http://www.quizballs.com`), and QuizZone (`https://www.quiz-zone.co.uk`). We remove duplicate entries and the remaining pairs are consolidated into a single source.

Each question is then run through two named entity recognition (NER) systems: TAGME (Ferragina and Scaiella, 2010) and FOFE NER (Xu et al., 2017). By combining the results of both

Components	Example 1	Example 2
Question [Answer]	Which 18th century author wrote Clarissa (or The History of a Young Lady), said to be the longest novel in the English language? [Samuel Richardson]	What is the correct name of the character voiced by Angela Lansbury in Beauty and The Beast? [Mrs Potts]
Subject (Freebase ID)	Clarissa (m.05s1st)	Angela Lansbury (m.0161h5)
Predicate	book.written-work.author	film.actor.dubbing-performances
Secondary Predicate	–	film.dubbing-performance.character
Object/Answer (ID)	Samuel Richardson (m.0hb27)	Mrs Potts (m.02vw823)

Table 1: Two typical examples to illustrate the data format of all matches in FreebaseQA.

systems, we create a list of possible subjects for each question. We use confidence thresholds of 0.2 and 0.9 for the respective systems to ensure that an adequate amount of entities are produced while avoiding the production of irrelevant results.

3.2 Freebase Matching

The matching starts by searching for all Freebase nodes with a name or alias matching each subject name. For each matched Freebase node (called a subject node), we search through all object nodes that are directly linked with the subject node. Then for each object node, we search through all of its names and aliases to see if one matches the answer to the question. Once a match is found, the subject node's Freebase ID, the predicate name, and the object node's Freebase ID are saved as a triple representing the question-answer pair. Note that one question-answer pair may generate several matched triples when multiple related predicates are found since each question may contain multiple entities and each subject node may link to an object node through different predicates.

However, this procedure becomes inefficient since there is an enormous number of object nodes to process for some popular subject nodes, such as United States (m.09c7w0), leading to a tremendous number of Freebase queries. Since we know the end point of the search, the answer to the question, this procedure is optimized by also starting from the answer and searching for all object nodes with a name or alias matching it. Then, the search concludes when the same object node is found from both starting points of the search. By using this two-way search method, we have accelerated the Freebase matching algorithm more than ten-fold.

3.3 Mediator Nodes

Freebase has been constructed with some special nodes called *mediator* nodes. A mediator is an intermediate node that connects a subject node with an object node. Since it itself is also considered a node, there are predicates from the subject to the mediator and from the mediator to the object. These mediator nodes are special as they do not have a name or alias associated with it, and only occurring in Freebase when there are multiple subjects and objects that are related through the mediator. When constructing the FreebaseQA data set, mediators are also accounted for. If the above search procedure reaches a mediator, a 2-hop matching strategy is conducted to search all nodes linked to this mediator. This captures a secondary predicate that bridges the subject to the answer through a mediator node. An example involving a mediator is described as Example 2 in Table 1.

3.4 Human Annotation

Since the matches found through the previously-explained algorithm are not guaranteed to be completely relevant to the question, human verification of the produced results is required to remove all possible false postitive matches. A group of 10 native English speakers are hired to label all of the collected matches. Each match is rated by the individuals as either "Completely Relevant", "Somewhat Relevant", or "Not Relevant". The choice of rating is dependent on the relevancy of the predicate to the question. If the predicate completely reflects the main idea asked by the question, the match is rated "Completely Relevant". If the predicate reflects part of the main idea of the question or is only somewhat related to it, the match is rated "Partially Relevant". Otherwise, the match is rated "Not Relevant". Compared with other

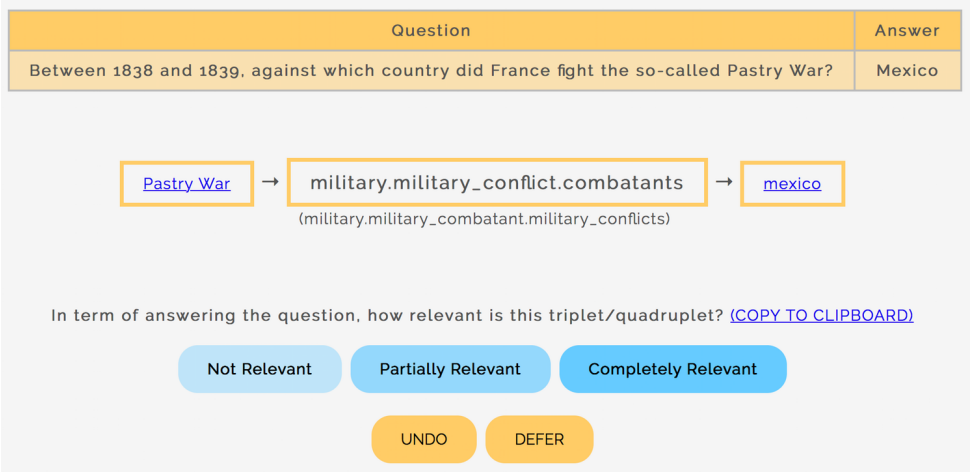

Figure 1: Human annotators use this website interface to label all automatically-generated matches, rating either Completely Relevant, Partially Relevant, or Not Relevant.

QA data collection tasks, human involvement in FreebaseQA is relatively light since each person only needs to make a one-out-of-three choice instead of composing a question or sentence from scratch. Therefore, using this method, we may significantly reduce the cost of QA data collection. As an illustration, the user interface for this data annotation procedure is shown in Figure 1.

In order to facilitate model training, the matches rated "Completely Relevant" are randomly chosen to populate the training, evaluation, and development sets of FreebaseQA. These sets are separated so that if there are multiple matches for a single question-answer pair, all of those matches will exist in only one of the sets. Moreover, the matches rated "Partially Relevant" are provided as a separate set, which may be useful for model training as well. The FreebaseQA data set is available for public use at `http://github.com/infinitecold/FreebaseQA`.

4 Results

We report the preliminary results of our statistical analysis on the collected FreebaseQA data set.

4.1 Collected Raw Question-Answer Pairs

The statistics of the originally collected question-answer pairs and the corresponding Freebase matches are summarized in Table 2. We see that with the exception of KnowQuiz, the number of matches in Freebase roughly equate the number of questions in each source. Among all the generated matches, 54,611 matches in total are kept as true positives by human annotators.

Source	Questions	Matches
TriviaQA	98,973	99,523
KnowQuiz	9,996	2,389
QuizBalls	15,370	17,856
QuizZone	7,686	7,289
Total	132,025	127,057

Table 2: A summary of the number of question-answer pairs from each source along with the number of matches generated from the above Freebase matching procedure.

4.2 FreebaseQA Statistics

The size of the FreebaseQA data set is compared to two similar QA data sets, WebQuestionsSP (WebQSP) (Yih et al., 2016) and SimpleQuestions (Bordes et al., 2015), in Table 3.

Data Set	train	dev	eval	Total
FreebaseQA	20,358	3,994	3,996	28,348
SimpleQuestions	75,910	10,845	21,687	108,442
WebQSP	3,098	-	1,639	4,737

Table 3: Total numbers of unique questions found in the subsets of each data set.

We see that FreebaseQA has a significantly larger size than WebQSP in number of unique questions, but it is about one quarter of SimpleQuestions in number of unique questions. Among these matches, FreebaseQA contains 28,348 unique questions in total, with 20,358, 3,994 and 3,996 in the train, dev and eval sets respectively.

However, another important factor to consider

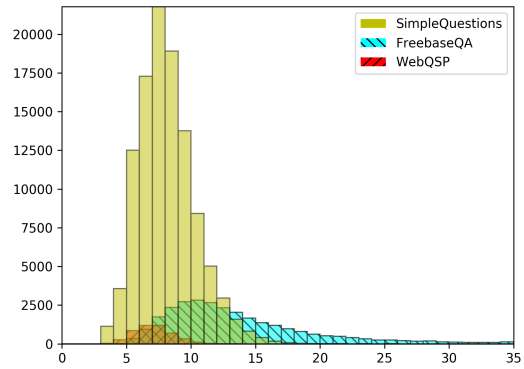

Figure 2: A histogram showing the spread of the length of the questions in each data set.

is the linguistic sophistication of the data. The sophistication of the linguistic structure of the questions in the FreebaseQA data set is compared to other similar data sets based on the average length, in number of words, of the questions. The histogram of question lengths of three data sets is shown in Figure 2. From the histogram, we see that the length of the questions in FreebaseQA extend much longer than the questions in SimpleQuestions or WebQSP (Yih et al., 2016). In fact, SimpleQuestions has an average length of 7.65 words per question and WebQSP has an average length of 6.62 words per question, while FreebaseQA has an average length of 13.35 words per question: approximately double the length of either data set.

4.3 Baseline Performance on FreebaseQA

Finally, we use FOFE-net (Zhang et al., 2015; Xu et al., 2017) to build a baseline KBQA system on FreebaseQA, which consists of subject detection, entity linking and relation detection in the pipeline. Our FOFE-net models are first compared with the popular hierarchical residual BiLSTM in (Yu et al., 2017) on two public data sets, such as SimpleQuestions and WebQSP. See (Wu et al., 2019) for more details on experimental settings and results. The comparison results are listed in Table 4.

As shown in Table 4, our baseline has achieved strong performance on the two public data sets but its final question answering accuracy has dropped significantly down to 37.0% on FreebaseQA. Obviously, FreebaseQA is a much more challenging KBQA task than both SimpleQuestions and WebQSP due to the fact that the questions in FreebaseQA are more complex in linguistic structure. Therefore, FreebaseQA may serve as an excellent testbed for more advanced KBQA techniques.

Data Set	BiLSTM (Yu et al., 2017)	FOFE-net (this work)
SimpleQuestions	77.0%	77.3%
WebQSP	63.0%	67.6%
FreebaseQA	-	**37.0%**

Table 4: Comparison of end-to-end QA accuracies on several KBQA data sets.

To facilitate the evaluation of the end-to-end question-answering pipeline on FreebaseQA, we have extracted a subset of Freebase, which contains all nodes and their corresponding predicates matching any entities in the FreebaseQA data set. This Freebase subset, also available at `http://github.com/infinitecold/FreebaseQA`, may be used to conduct end-to-end QA experiments to compare with our performance results in Table 4.

5 Conclusion

This paper presents a new data set, *FreebaseQA*, for open-domain factoid QA over structured knowledge bases. FreebaseQA has a size of over 54K matches, significantly larger than WebQSP and linguistically more sophisticated than SimpleQuestions. Our baseline QA results have also shown that FreebaseQA is a much more difficult KBQA task than either WebQSP or SimpleQuestions. Therefore, FreebaseQA may be an invaluable asset to the investigation of more advanced machine learning methods for factoid KBQA problems. Furthermore, the use of this data set is not only limited to factoid question answering, but several other applications can also be approached with this data set, including reading comprehension, natural language-based search, and the quantification of natural language understanding.

Acknowledgments

This work is partially supported by a research donation from iFLYTEK Co., Ltd., Hefei, China, and a discovery grant from Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1533–1544. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. Association for Computing Machinery.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Weston Jason. 2015. Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1625–1628. Association for Computing Machinery.

Google. 2017. Freebase data dumps. `https://developers.google.com/freebase/data`.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611. Association for Computational Linguistics.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. arXiv preprint arXiv:1607.06275.

Chen Liang, Jonathan Berant, Quoc Le, Kenneth D. Forbus, and Ni Lao. 2017. Neural symbolic machines: Learning semantic parsers on Freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23–33, Vancouver, Canada. Association for Computational Linguistics.

Lili Mou, Zhengdong Lu, Hang Li, and Zhi Jin. 2017. Coupling distributed and symbolic execution for natural language queries. In *Proceedings of the 34th International Conference on Machine Learning (ICML-17)*, pages 2518–2526.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading COmprehension dataset. arXiv preprint arXiv:1611.09268.

Michael Petrochuk and Luke Zettlemoyer. 2018. SimpleQuestions nearly solved: A new upper-bound and baseline approach. arXiv preprint arXiv:1804.08798.

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.

Iulian Vlad Serban, Alberto Garca-Durn, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. arXiv preprint arXiv:1603.06807.

Dekun Wu, Nana Nosirova, Hui Jiang, and Mingbin Xu. 2019. A general FOFE-net framework for simple and effective question answering over knowledge bases. arXiv preprint arXiv:1903.12356.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1237–1247. Association for Computational Linguistics.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206. Association for Computational Linguistics.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 571–581. Association for Computational Linguistics.

ShiLiang Zhang, Hui Jiang, MingBin Xu, JunFeng Hou, and LiRong Dai. 2015. The fixed-size ordinally-forgetting encoding method for neural network language models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 495–500. Association for Computational Linguistics.