Literature Survey for Knowledge Graphs in Question Answering Applications using Wikipedia Mined Data

Kriti Gupta, Amrit Gupta

Abstract

This review paper highlights the various research work undertaken on the subject of use of knowledge graphs for question answering systems, extracting Wikipedia and mining lexical semantic data from it and other related work in combining these. As wikipedia is a huge source of information classified into various categories, subjects and topics it can be a very useful source for answering questions and other natural language processing applications. In this paper we have highlighted concepts such as web questions, Simple questions, constraint based question answering, semantic approach to question answering, extracting lexical semantic data from Wikipedia, machine learning techniques used to answer open domain questions and factoids and knowledge graph based question answering.

Keywords: knowledge graph, question answering systems

1. Introduction

The extraction of knowledge from different web based resources, such as Wikipedia is an important topic for research as the web is a cesspool of information and navigation is typically a time consuming operation. Question answering systems provide a much better option for users who do not wish to wade through complete documents in search for information. A lot of research has been conducted on these systems in the past and this paper aims to collate the insights from different studies regarding question answering systems and extract the gaps which can then be filled in further studies. The first studies on QA systems came up with knowledge based datasets such as WebQuestions and SimpleQuestions. These were very simple in nature and thus higher level queries were handled in multiple ways such as creation of a more complex base, changing the approach to semantic based systems, and knowledge graph embedding. Wikipedia itself was the subject of much research, starting from extraction of lexical knowledge to creation of neural nets for the better parsing of natural language.

2. Reviews

2.1 WebQuestions And SimpleQuestions

WebQuestions and SimpleQuestions are two datasets created explicitly for the development and improvement of knowledge based question answering systems. These, along with QALD, are some of the most important benchmarks in question answering systems. WebQuestions contains 5810 questions. They can be answered using one reified statement with potentially some constraints like type constraints or temporal constraints. SimpleQuestions contains 108.442 questions which can be answered using one triple pattern. Both of these are used for answering single-relation, factoid questions on finite data. These are fixed datasets and thus cannot be used to resolve questions on newer and ever-changing data which is what is required for our use-case: question answering based on wikipedia documents.

2.2 Constraint based Question Answering

Since WebQuestions and SimpleQuestions could not handle multi-constraint questions, a new knowledge based Dataset was devised called ComplexQuestions that could handle the same. It catered to a variety of constraints such as multi-entity constraints, type constraints, temporal constraints, ordinal constraints and aggregation constraints. This new dataset offered better semantic structure and availability of access for multiple connections. A multi-entity query graph could be constructed to denote the knowledge captured. While a step up from WebQuestions and SimpleQuestions, ComplexQuestions is still a fixed database and thus cannot be directly used for our applications. Our data is dynamic and not extremely large. Thus the complexity of overlaying this database over the application is unnecessary.

2.3 Semantic Approach to Question Answering

The question answering system in the semantic approach consists of the following 3 main steps:

(1) Triple pattern extraction: Candidate RDF triples are extracted from natural language questions using the dependency graph and POS tags of the questions.

(2) Entity and property extraction: Using RDF triples from the previous step, all subject, predicate and objects are mapped to DBpedia entities, classes or properties.

(3) Answer extraction: Candidate triples are queried over DBPedia and all the answers matching the expected type of a question are ranked and the top result is given as an answer.

2.4 Extracting Lexical Semantic knowledge from the web

This paper brings up the usability and utility of Wikipedia and Wiktionary in the field of NLP. This paper addresses the lack of suitable programmatic access mechanisms to the knowledge stored in these large semantic knowledge bases. They have presented two application programming interfaces for Wikipedia and Wiktionary which are especially designed for mining the rich lexical semantic information dispersed in the knowledge bases, and provide efficient and structured access to the available knowledge. The paper greatly explains about collaborative knowledge bases and its differences and similarities with Linguistic knowledge bases.

2.5 Validating answers and using categories in wikipedia effectively

In this paper they have investigated the use of Wikipedia, the open domain encyclopedia, for the Question Answering task. Previous works considered Wikipedia as a resource where to look for the answers to the questions. They have focused on some different aspects of the problem, such as the validation of the answers as returned by their Question Answering System and on the use of Wikipedia "categories" in order to determine a set of patterns that should fit with the expected answer. Validation consists in, given a possible answer, saying whether it is the right one or not. The possibility to exploit the categories of Wikipedia was not considered until then. They performed their experiments using the Spanish version of Wikipedia, with the set of questions of the last CLEF Spanish monolingual exercise. Results showed that Wikipedia is a potentially useful resource for the Question Answering task.

2.6 Answering open domain questions using TF-IDF & RCNN from Wikipedia data

This paper proposes to tackle open domain question answering using Wikipedia as the unique knowledge source: the answer to any factoid question is a text span in a Wikipedia article. This task of machine reading at scale combines the challenges of document retrieval (finding the relevant articles) with that of machine comprehension of text (identifying the answer spans from those articles). The approach combines a search component based on bigram hashing and TF-IDF matching with a multi-layer recurrent neural network model trained to detect answers in Wikipedia paragraphs. They have experimented on multiple existing QA datasets indicate that:

(1) both modules are highly competitive with respect to existing counterparts
(2) multitask learning using distant supervision on their combination is an effective complete system on this challenging task

2.7 Knowledge graph Embedding based Question Answering

Question answering over knowledge graph aims to use facts in the knowledge graph to answer natural language questions. It helps end users more efficiently and more easily access the substantial and valuable knowledge in the knowledge graph, without knowing its data structures. Question answering knowledge graphs is a nontrivial problem since capturing the semantic meaning of natural language is difficult for a machine.. Knowledge graph embedding targets at learning a low-dimensional vector representation for each predicate/entity in a knowledge graph, such that the original relations are well preserved in the vectors. These learned vector representations could be employed to complete a variety of downstream applications efficiently. Examples include knowledge graph completion, recommender systems, and relation extraction.

3. Conclusion

The problem of building question answering systems is not a new one and thus much research has already been conducted on it. Despite this, there are few satisfactory QA systems today and thus there is clearly a gap in the research and user expectations. There is also little research done on non-fixed datasets, i.e, dynamic datasets, such as wikipedia, since the training of neural networks and such is a resource and time intensive process. WebQuestions and SimpleQuestions are much too simple for our application domain and the use of neural nets takes too much time. Therefore the best approach that can be taken is a combination of semantic and lexical analysis with knowledge graphs.

References

[1] Diefenbach, D., Lopez, V., Singh, K. et al. Core techniques of question answering systems over knowledge bases: a survey. Knowl Inf Syst 55, 529–569 (2018).

[2] Bao, Junwei, et al. "Constraint-based question answering with knowledge graph." Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016.

[3] Diefenbach, Dennis, et al. "Question answering benchmarks for wikidata." 2017.

[4] Zesch, Torsten, Christof Müller, and Iryna Gurevych. "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary." LREC. Vol. 8. No. 2008. 2008.

[5] Buscaldi, Davide, and Paolo Rosso. "Mining knowledge from wikipedia for the question answering task." Proceedings of the international conference on language resources and evaluation. 2006.

[6] Chen, Danqi, et al. "Reading wikipedia to answer open-domain questions." arXiv preprint arXiv:1704.00051 (2017).

[7] Hakimov, Sherzod, et al. "Semantic question answering system over linked data using relational patterns." *Proceedings of the Joint EDBT/ICDT 2013 Workshops*. 2013.

[8] Huang, Xiao, et al. "Knowledge graph embedding based question answering." Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019.