Design of KBQA System

Yidong Zhao, Hongru Guo

June 2020

1 Introduction

1.1 Something about KBQA

KBQA is a question answering system based on knowledge base, which brings questions into the prepared knowledge base to seek answers. Specifically, from the perspective of application field, knowledge base QA can be divided into Open Domain Knowledge QA, such as encyclopedia knowledge QA, and specific domain knowledge QA, such as financial field, medical field, religious field, etc., which serve our daily life in the form of customer service robot, education / examination robot or search engine. Nowadays there are many mature KBQA systems such as:Watson of IBM and DeepQA of Microsoft. However they are Open Domain Knowledge QA, they can answer any question in general domain. However, as for specific domain knowledge QA, there are also some perfect experiments, such as KBQA for Medical Science used for choose correct medcine, and KBQA for poems, use for ask question on poem domain.

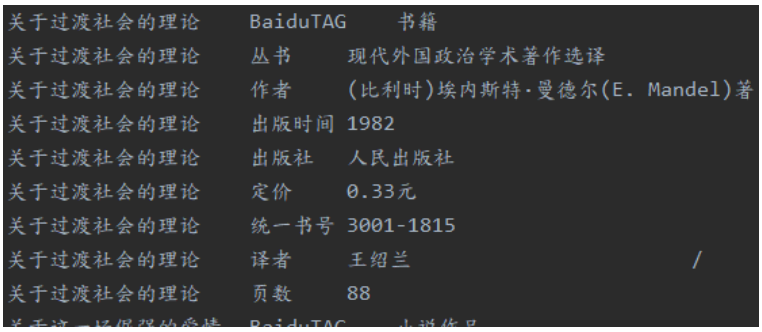
1.2 Our Work

Our KBQA system is also a specific domain knowledge QA, which is base on knowledge base of CNDBpedia Dump and used for asking question of book's domain. What we mainly did is: design a model of KBQA system asking books, which need us to understand the problem, then find the correct answer.

2 Data Set

Our data set is based on CNDBpedia Dump, which is design by FudanUniversity. It's a professional data set, mainly extracts information from the plain text pages of Chinese encyclopedia websites (such as Baidu Encyclopedia, Interactive Encyclopedia, Chinese Wikipedia, etc.), and finally forms high-quality structured data for machine and human use after filtering, fusion, inference and other operations.

Our reason using this data set is: Firstly, this data set is orderly, under the format of entity-relation-entity. The advantage of this structure is that it is more convenient to retrieve and classify, and it is more convenient to operate, it does not need to crawl the data in advance and it is faster to retrieve. Secondly, this data set is more complete. Under our obse. Through our observation, we found that this data set is very comprehensive, and the book related content accounts for a large proportion, which can be used to replace the crawled special book database.
The data base we used is mysql. We put the data in the database in the following format and add indexes to the entities and relationships before the relationship. The reason is that the size of the entity after the relationship is long, which exceeds the maximum index size, so the entity after the relationship cannot be set as an index.



(a) Structure of Data

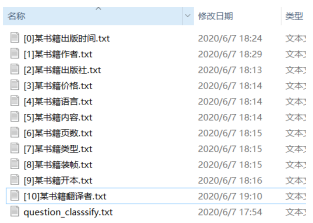
(b) Saved in Database

3 Core Module

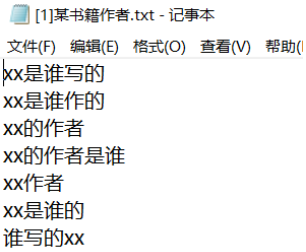
Our model is divide into three part:

3.1 Question Classification Module

First is question classification module. The function of this module is to build a question tree, that is, to divide questions into several categories first, and then come up with corresponding questions under each category, and then build a good question tree. These problems are all obtained through experience, which is what we think by ourselves. The graph is as following. And we also create a index file to locating the file and end of each index is the key words of each problem, which is used in result searching.



(c) Model1



(d) Model1

(e) Model1

3.2 Question Analysis Module

Second model is question analysis module. This model is the core part of the total algorithm. The function of this part is to understand what the questioner ask, is the head of the model. Our method is participle input question (using jieba participle). Then starting from the question [0], compared each question in question tree (has been participled) in turn. (before the participle, "XX" is removed), then the matching degree p is obtained. The algorithm is as following. In this algorithm we arise the weight of the key words, which means if we obtain the key words we rise the value of P.
What more we need to find the main entity the questioner ask. Because we don't have the entity data base to search, so we have to find another way. Our method is to divide the question into multiple part, and for each part find whether it is book by searching in the data base. And using the book entity.

$$p = \lambda * 1(exist\_key\_word) + (1-\lambda) * (\sum_{each\_word} 1(exist\_this\_word))/(N-1)$$

(f) Algorithm

3.3 Result Selection Module

Third model is result selection module. In this model, we using pymysql searching in the data base. We using a "for" loop to search because we may have multiple result.

```
import handle_file
def get_answer(keywords):
    relations=handle_file.getrelations(keywords[1])
    db = pymysql.connect("localhost", "root", "123456", "fdudb", charset='utf8')
    cursor = db.cursor()
    for relation in relations:
        sql = 'SELECT entityb FROM fdudb WHERE entitya="%s" AND relation="%s"'.format(keywords[0],relation)
        cursor.execute(sql)
        res = cursor.fetchall()
```

(g) Result Selection

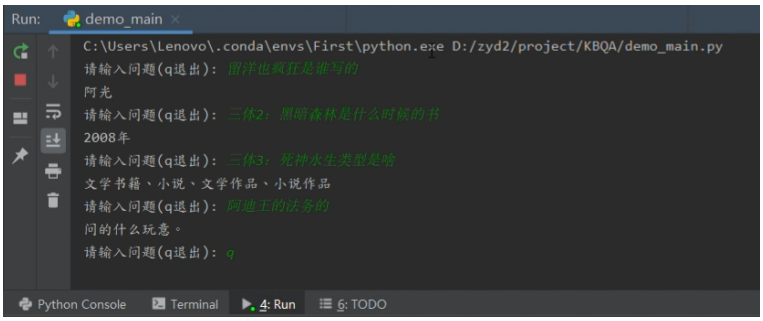
4 Experiment Results

Our experiment is in figure Result.

5 Conclusion

5.1 Our Deficiency

Firstly our project has some problems. Because only using word matching, we don't have very highly semantic comprehension ability. So when asking something like book "author" will arise problems. And also our problem can't handle



(h) Result

extremely colloquial situation. What's more when asking some books using not correct name will arise problems.

5.2 Feature Work

We also want to improve a lot of directions, such as using self-update algorithm. When there are multiple possible subjects in the search, we need to ask what the subject is before outputting the results, and then save his question style in the question set of such questions. Similarly, the types of problems can also be self-renewal. Another way is to change the data set, save the data of books, or crawl the data of a book.

6 Citation

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 428-438. Springer, Cham, 2017.

7 Division of labor

Yidong Zhao:
report: section Core Module, Experiment Results, Conclusion.
Work: think of model, Model implementation, Writing PPT and presentation;

Hongru Guo:
report: section Introduction, Data Set.
Work: think of model, Model implementation of data part, Writing PPT and presentation;