An Approach for the Chinese Question-answer System based on Document

Benyou Wang[1], Jiabing Niu[1], Liqun Ma[1], Yuhua Zhang[1], Lipeng Zhang[1], and Peng Zhang[1]

Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, Tianjin, China

Abstract. Question Answering system has gradually become a new trend within the field of information retrieval and NLP. It outperforms the conventional search engines, for the system is able to answer users questions automatically and accurately. Question Answering system based on English corpus has developed rapidly, whereas the Chinese corpus based Question Answering system still has some problems remains to be solved. Thus, developing a new Question Answering model, which is characterized by dealing with features of Chinese corpus is extemely essentail. Different to the current deep learning model, our model uses the semantic and syntactic information in Chinese corpus and bases on the linearity of Chinese texts. Finally, our model turns out to perform better than other methods through experiments.. . .

Keywords: Question Answer, graph theory, Hamilton cycles, linearity, Chinese corpus

1 Introduction

Quesion Answering is gradually becoming a more and more significant research area with the development of Natural Language Processing and Information Retrieval techniques. Conventional search engines, such as bing, Google and Baidu, are keywords based systems which are able to return a large number of results containing hyperlinks related to keywords in users' queries. However, users usually have to browse dozens of results to find their target answers if the first few results couldn't meet their needs. Therefore, question answering system is designed to answer users' short question sentences as well as combined phrases directly and accurately, for only one exact answer will be returned by the question anwsering system, which outperforms the conventinal search engines to some degree. Meanwhile, there are still some difficulties remain to be solved in the Chinese corpus based question answering system compared with the English corpus based one. For example, QA

This paper elaborate our methods and experiments for the Open Domain Questin Answering shared sub-task of Document-based QA task in NLPCC 2016. The aim of the task is to build a system which can predict answers to each given question. The target answers are only supposed to be selected from the question's

given document contains a set of answer sentences. Our work is to train necessary models to generate answers which are highly related to the given questions. And the results will finally be evaluated by the evaluation metrics to figure out the performance of our system. For example, for the given question ?in training set, the correct answer 6364879.43.15 will be labled $/\widehat{1}$ and the rest answer sentences to the question will all be labled $/\widehat{0}$. Our model should generate a set of relevance scores between the question and each answer sentence including the correct answer in the test set. The evaluation toolkit which has test sets labeled golden answer annotations to questions will rank all answer sentences according to these scores. And finally give the MRR and MAP value of our model as the evaluation metrics.

We propose a method which integrates various features extracted from the question and answer sentences to form a model, which makes full use of the semantic and syntactic information as well as the linearity of Chinese corpus. . And our model outperforms other models including deep learning model according to the experimental results.

The structure of this paper can be listed as follows. Section 2 introduces some work related to the question answering system. Section 3 describes the methods including features and model which we use to complete the shared task. Section 4 presents the experiments and results we get. Section 5 are discussions and conclutions.

2 Related Work

just test a cite [1]

3 Methods

We adopt a classical process for a competition task, with the successive steps from data exploration (in Sec 3.1), data preprocess (in Sec 3.2), feature extraction (in Sec 3.3) to model selection (in Sec 3.4).

3.1 Data Exploration

Word Overlap and Character Overlap It is considered that the more keywords in questions are matched with those in the answer sentences, the more likely that the answer is the correct one.We have found that there are some relationships between the overlapped characters or words and the occurrence frequency of the correct answer after analysing the given corpus.

We first segment the question and answer sentences into a series of characters and then count the total overlapped chatacers between the question and answer characters.As what has been shown in Fig. 1,it is concluded that there is a linear dependence between the correct answers frequency and overlapped characters.The same methods are also applied to the overlapped words and Fig. 2 illustrates that the more words are overlapped,the more likely that the answer is the correct one.

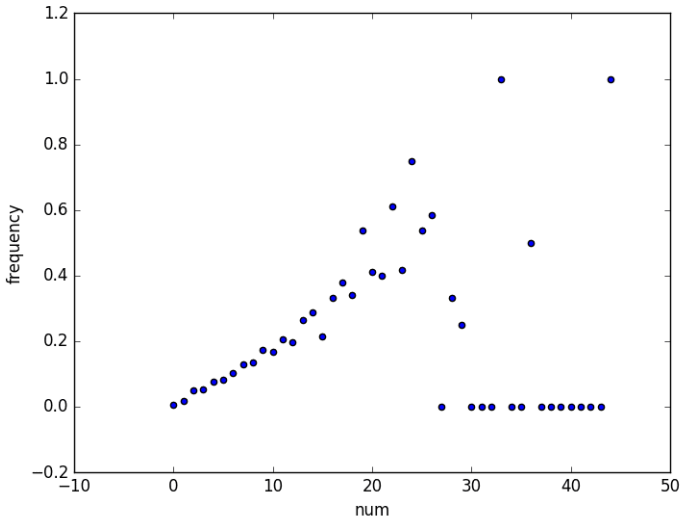

Fig. 1. x-axis means the number of overlapped characters in both question and answer sentences.y-axis refers to the probable occurrence frequency of the target answer. Data are dispersed in the range between 22 and 50 because the samples are not enough.

Position Message in Overlapped Words In Chinese grammer, different grammatical elements are usually distributed in different position in a sentence. For example, in a classic question, the entities concerned usually turn up in the front of a sentence, which follows the interrogatives like ,.People have various speaking habits when they organize a sentence in Chinese.For instance, premises are more inclined to appear first in one sentence.So it is reasonable to consider keywords position message when we are trying to get the degree of how question and answer matches. We try to give different weights to keywords in different positions in a sentence.Meanwhile, weights may also be concerned with words idf values, which means the discrimination of words.

Firstly, we segment the question and answer sentences into characters (words) and then count the overlapped characters(words) and their position in both questions and answers.It is obvious that the overlapped characters(words) which often appear in the back of one sentence tends to be more important for us to find out the correct answer among answer text sets. We can also conclude from the two figures above that keywords or characters are more likely to appear in the back of the question sentences. Thus, we should give them higher weight than the rest. Unforturnatly, word-overlap based method is still based on the hypothesis of word independence. Synonym rewriting based methods and translation model could model words which are extremely closed in meaning to some extent. To enumerate a rewriting templet(translate templet) of high quality and maturity is never a easy task. Current embedding methods, to some degree, model this kind

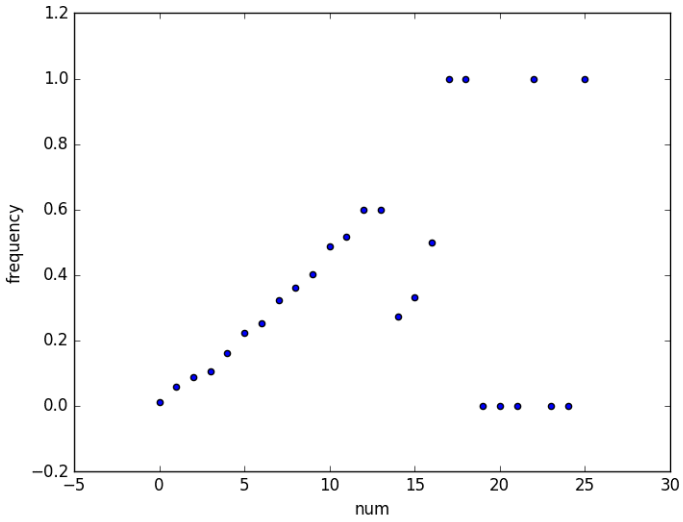

Fig. 2. x-axis means the number of overlapped words in both question and answer sentences. y-axis refers to the probable occurrence frequency of the target answer. Data scatters after the number of 14 for the lack of words samples.

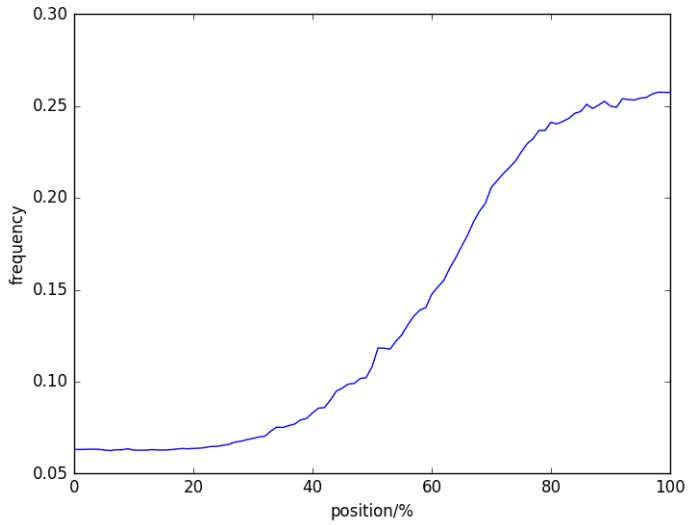

Fig. 3. x-axis refers to the position of the overlapped characters in question sentences. x=0 means that the overlapped character is on the front of a sentence.x=100% means the overlapped character is on the back of a sentence.y-axis means the occurrence frenquency of correct answers.

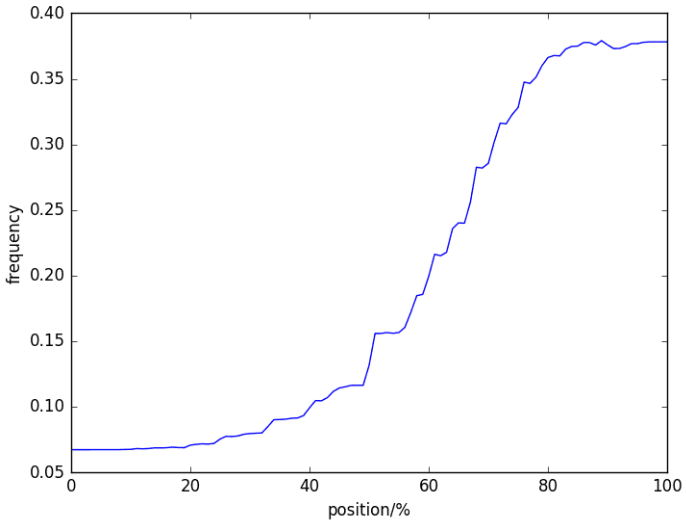

Fig. 4. Similar to 3, this figure shows the relationship between the position of the overlapped words in question sentences and the occurrence frenquency of correct answers.

of semantic similarity from semantic space of high dimensional, which will be analysed in Sec 3.3. In Chinese, word group is composed of more Fined-grained characters. And word based overlap is a crude method to consider the correlation between words, which can be combined with other methods.

3.2 Data Preprocessing

We uses the pynlpir package [1] to segment both the question sentences and the answer texts into word groups and we base our next work on these words.We then omit some of the stop words like in those word groups according to the stop words list(http://).It is necessary to cut out several meaningless words since it is of great help to extracting the words overlap feature.Finally, we get a processed dataset of words.

3.3 Feature Extraction

Questions and Answers Type Question sentences type is usually a kind of vital signal.It is a prerequisite but not sufficient condition when distinguishing whether the given sentence is a correct answer or not. For example, for the question ,when the name entity related to person appears in one of the answer texts,it is probably that this sentence is the exact correct answer.But if little

[1] http://github.com/tsroten

information is mentioned in the sentence,it is certain that the sentence will never be the target answer.Therefore, the type information helps to classify questions and answers as well as find the target answers.

The answers to questions can be divided into 5 types() and 3 of them are concerned with name entity, which are Person, Place, Organization. The others are Time and Number. In fact, a learning based way of finding out which type the question and answer belongs to could also make sense.However, we adopt the methods of extracting name entity to define the types of Person, Place and Organization and use model to find out the types of Time and Number because of the lack of supervised training and testing samples for learning. It turns out that our methods could cover more than 80% cases and receive better results.

We adopt the LTP online API [2] for Named Entity Recognition (NER). Meanwhile, we use keywords templets to distinguish the types of Time and Number.

In coding, two methods are adopted. The first one is Dummp Number. The 5 types are viewed as 5 kinds of vectors. For each type, we use one-dimensional vector to indicate whether Question and Answer are matched or not.At the same time, or operation is made among the 5 types and the result is viewed as an one-dimensional vector for the reason that the result of each type is too sparse to count. The second method is that the answer texts which are well matched by keywords in the question are marked as 1 and those which are not matched are marked as 0.

Overlap As QA pair match and Correlation have been mentioned in Sec 3.1, namely,the more words are overlapped in both question and answer sentences, the more likely that the answer is the target one. In these features, the removation of stop words seems to be vital due to the ignorance of the meaningless words. For Chinese text, it's hard for us to find the dictionay which can contains all the near-synonym pairs. An alternative approch is to use the charcter-based (or) metric due to the fact that many near-synonym pairs shares the same charcter in Chinses.

Other Conventional Methods Edit distance is usually used to measure the length of similar strings in English. While Jaccard index gives the similarity of two sets.And a sentence is often considered as a set of words. The length of the answer texts is also a significant feature.

Embedding-based approach We use embedding model as our main model. Embedding means to embed words into a high dimensional semantic space, which makes it easier to find the relationship between words. Sentence is simply considered to have been lapped by words linearly. As what has been mentioned in Sec 5, different words can contribute different weights for the whole meaning of a sentence, which depends on their position, semantic structure and idf.

[2] http://www.ltp-cloud.com/

Besides these linear combination of the inside word embedding, we have build a CNN model to test how much the question-answer pair matches.

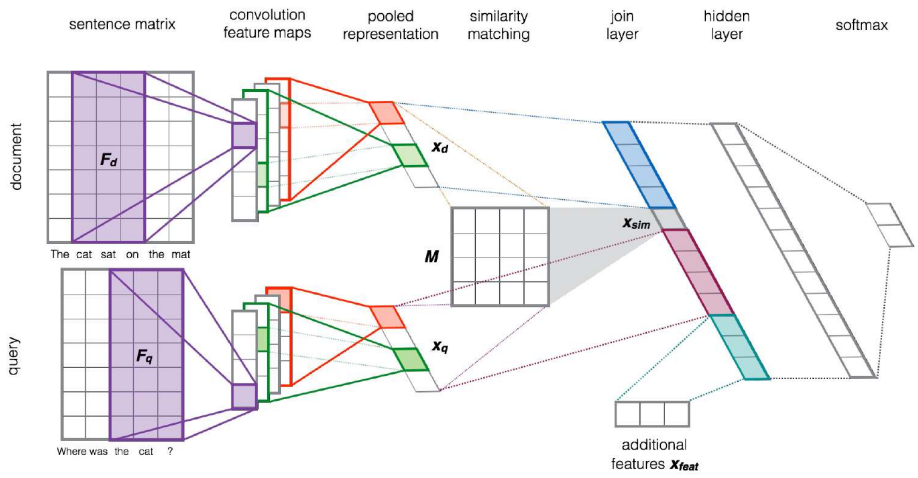

Fig. 5. The picture need to be repainted

3.4 Model Selection

We have presented various fundamental features in last chapter,which will directly effect the degree of how question and answer matches.We adopt a linear regression model to integrate those features.While some integrated learning methods like Ada Boost and bagging models are also used.

4 Experimental results

4.1 Dataset

The provided datasets [3] of the document-based quetion answer task contains a training data which have the label for the ground truth and a testing data which do not have the correct label. In the training dataset, each question have many candidate answers (mean :20.73, var: 112.29, max 30:, min:1)

[3] http://tcci.ccf.org.cn/conference/2016/pages/page05_evadata.html

4.2 Evaluation Metrics

Our Question Answering system will be evaluated by MRP and MAP. Mean Reciprocal Rank(MRR) mainly indicates that whether the recall results good or not depend on the rank of the correct answer, namely the higher the correct answer ranked, the better the results are.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|} Q|\frac{1}{rank_i} \tag{1}$$

$|Q|$ stands for the total number of questions in the evaluation set. For the i^{th} question Q_i, $rank_i$ represents the position of the first correct answer in the generated answer set C_i. $\frac{1}{rank_i}$ equals 0, if C_i doesn't have the correspond answer with the golden answers A_i for Q_i. For example, if there are 4 queries in the test set. Suppose the correct answers for the first 3 queries are ranked 3,1,5 respectively and there is no answer for the last query.

$$MRR = (\frac{1}{3} + \frac{1}{1} + \frac{1}{5} + 0)\frac{1}{4} = 0.383$$

Another evaluation metric is Mean Average Precision(MAP), which can be defined as follows.

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|} Q|AveP(C_i, A_i)\ AveP(C, A) = \frac{\sum_{k=1}^{n}(P(k)\cdot rel(k))}{min(m,n)} \tag{2}$$

MAP mainly refers to the average precision of the results. If the correct anwser ranks high in the retrieved answer set, the value of MAP will be high accordingly.m denotes the number of correct answer sentences and n the number of retrieved answer sentences. $P(k)$ is the precision at cut-off k, which is the rank in the sequence of retrieved answer sentences. $rel(k)$ equals 1 if the item at rank k is an answer sentence, and 0 if not.

4.3 Results

5 Discussion and Conclution

chinese [2]

References

1. A. Severyn and A. Moschitti, "Learning to rank short text pairs with convolutional deep neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 2015, pp. 373–382.
2. Y. Li, W. Li, F. Sun, and S. Li, "Component-enhanced chinese character embeddings," *Computer Science*, 2015.