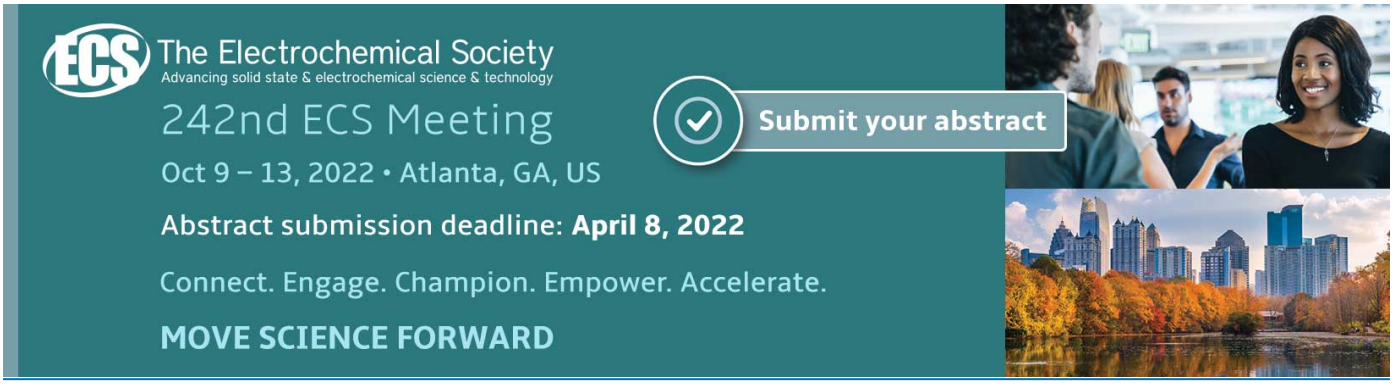
Question answering system based on tourism knowledge graph



Question answering system based on tourism knowledge graph

Yuan Sui[1*]

[1]Department of information science and engineering, Shandong Normal University, Jinan, Shandong, 250300, China

[*]Corresponding author's e-mail: 201911010105@stu.sdnu.edu.cn

Abstract. Nowadays tourism information services only provide users with massive and fragmented information returned by independent network search which makes users often need to spend a lot of time and energy to find what they really want from the massive data. As a result, route designing is very complicated. In view of this situation, this study builds a tourism knowledge graph based on neo4j and constructs a question answering system (QA). Also, we carry out the model and system performance evaluation, trying to improve the user satisfaction with query experience. According to the structure of question answering system (QA), this research designed and implemented named entity recognition (NER) model based on Bert-BiLSTM-CRF and matching reasoning model based on templates. With the above methods, natural language questions were successfully transformed into cypher query statements recognizable in graph database, and the corresponding answers will be captured and returned from tourism knowledge graph. According to the experiment, the method of Bert-BiLSTM-CRF obtains the state of art and QA system performs quickly and efficiently. For the purpose that artificial intelligence helps the development of tourism industry, this study has a certain significance.

1. Introduction

Question answering system (QA)[1] is a research direction in the field of artificial intelligence and natural language processing, which attracts much attention and has broad development prospects. It can transform the questions described by users' natural language into machine recognizable query sentences, and return the information with accurate and concise natural language [2]. As an important reference scenario of QA system, a perfect tourism Knowledge Based QA system has the ability to help people quickly obtain tourism related information, understand the relevant culture and characteristic tourism resources of tourism destination, and provide technical support for users to specify personalized tourism scheme through typing into the query sentences. However, the current situation of query service for the tourism industry is not satisfactory: On one hand, tourists' demand for cultural knowledge and tourism information is increasing; On the other hand, due to insufficient investment in the construction of tourism knowledge automatic question answering system, the current consulting service is still in the stage of users searching fragmented information on the Internet alone, far from meeting the needs of users.

At present, the question answering system based on knowledge graph has been fully studied and applied in many different fields, and has improved the retrieval efficiency of domain knowledge. In the fsield of finance, Chongyu Zhang (2019) [14] proposed an automatic question answering method for clinical medical knowledge graph. Relying on the clinical medical knowledge graph, the expectation

was generated through the data cold start mechanism and Bi-LSTM+CRF algorithm, which achieved good results; Siyu Wang et al. (2020) [15] proposed an online question answering system based on online knowledge graph and rule reasoning. By constructing attribute attention network based on LSTM and introducing ontology reasoning rules, the knowledge graph can dynamically generate a large number of triples, so that more questions can be answered under the same data. Experimental results show that the system has good performance. In the field of movie original sound, Dongjin Huang et al. (2020) [16] proposed an entity recognition method based on rules and dictionaries and a classification algorithm based on Bert+CNN. The experimental results show that the constructed intelligent system for movie original sound can get accurate feedback of the answers in real time.

2. System framework

The question answering system based on tourism knowledge graph designed in this study firstly needs to preprocess the natural language questions input from the users, then, extracts the entities in the questions through the Bert-BiLSTM+CRF model [3]. Then match and analysis the present query template and the questions according to the system, and divide the categories of the questions through the improved NB algorithm, so as to find the corresponding query template. The obtained entity and entity relationship will be embedded in cypher statement construction. Based on the above steps, query statements can be used to obtain the answers from the knowledge graph and return to the users. Considering that the actual product design generally has the characteristics of modularization and structure, the system is divided into three parts and eight models according to the functions. The system framework is as follows in Fig 1.

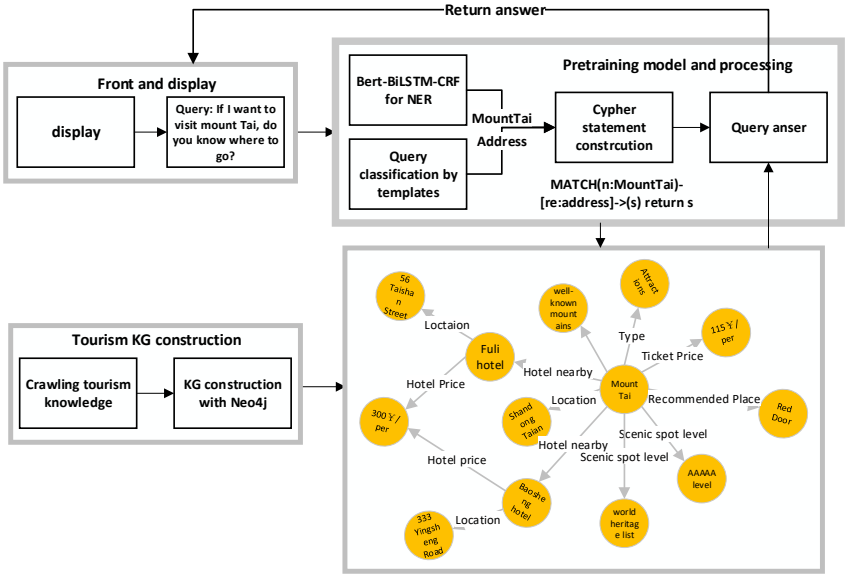


Figure 1. Framework of the QA system based on tourism knowledge graph

3. QA System

This research mainly uses Bert (the bidirectional encoder representation based on transformer proposed by Jacob et al.) to fully use the contextual information and directly process the original data. And use BiLSTM, a bi-directional long-term and short-term memory network proposed by Hochreiter and Schmidhub in 2014)[5][6] to extract the semantic entities of text in depth by adjusting the input gate, the output gate and the forget gate.

Considering that although BiLSTM model can obtain word vector representation sequence well, its output results often have the problem of scattered word annotation results, a typical discriminant model CRF layer proposed by Lafferty et al in 2001 is added to decode the output results of BiLSTM model and optimize the annotation sequence. Based on the above steps, the named entity recognition task is realized.

Also, we established a template library for relationship matching and divided the categories of the questions through the improved NB algorithm [7], so as to find the corresponding query template. The obtained entities and relationships will be converted to Cypher sentences.

3.1. Bert-Pretreatment model

The core function of the Bert preprocessing model is to represent the input natural language corpus. Each word in the text is represented by multi-layer embedding and a series of numerical transformations. Finally, the final semantic representation of each word will be output. The key part of Bert model is to process the text based on attention mechanism instead of RNN (recurrent neural network). Its basic principle is to calculate the relevance and importance between each word and all words in the sentence. The expression of each word can be obtained based on the relevance of context as follows [8]:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \tag{1}$$

Where Q,K,V are the input word vectors, d_K is the dimension of the input word vector. The input of Bert is represented by the addition of the corresponding word vectors: token embeddings, segment embeddings and position embeddings. Compared with other language models, the pre training language model of Bert can make full use of the information on both sides of the word and besides, it can obtain the best distributed representation of the word.

3.2. Bi-LSTM and CRF model

Hochreiter et al. proposed using LSTM model to introduce memory units and gate restriction, which have the ability to realize long and short distance information and thus it can solve the problem of gradient disappearance.[9] However, there's still some gap needed to be fixed. In order to improve the efficiency, Graves et al. proposed methods to upgrade the door restrictions. In this paper, we use the Graves's proposed gate restriction. The specific calculation process of LSTM hidden layer output representation is as follows:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{3}$$

$$c_t = i_t \times \tilde{c}_t + f_i \times c_{t-1}, h_t = o_t \times \tanh(c_t) \tag{4}$$

where W is the connection between the two layers ($\mathrm{W_{xi}}$ is the weight matrix from the input layer to the hidden layer; b represents the offset vector (b_i is the offset vector of the input gate from the hidden layer; c_t represents the memory unit, $\tilde{c}_t$ represents the state at time t and tanh represent two different neuron activation functions; And i_t, f_t and $\mathrm{o_t}$ represents input gate, forget gate and output gate respectively.

Since BiLSTM does not consider the dependence between sequences, its output often contains meaningless characters.[10] However, the sentence-level log likelihood function (CRF) proposed by collebert et al. can reasonably consider the dependency relationship between tag sequences, and can better predict tags, so as to obtain the global information of tag sequences. The combination of BiLSTM and CRF can ensure that the final recognition result is correct.

For the input sequence $\mathrm{x} = (x_1, x_2, ..., x_n)$, it is supposed that P is the score matrix obtained from the output of Bi-LSTM network. And for the predict sequence $\mathrm{y} = (y_1, y_2, ..., y_n)$, we assume that the total score is: $\mathrm{S}(x,y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}$, where A represents the transition matrix score from tag i to tag j .The probability of softmax generating sequence y on all possible tag-sequences is as follows:

$$p(y \mid x) = \frac{(e^{s(x,y)})}{(\sum_{\tilde{y} \in Y_x} e^{s(x,\tilde{y})})} \tag{5}$$

In the training process, the logarithm probability of the correct tag sequence is maximized:

$$log(p(y \mid x)) = s(x,y) - log(\sum_{\tilde{y} \in Y_x} e^{s(x,\tilde{y})}) = s(x,y) - log \sum_{\tilde{y} \in Y_x} (x,\tilde{y}) \tag{6}$$

Where Y_x represents all possible tag sequences of input sequence x. By maximizing the logarithm concept, we encourage the network to generate effective output tag sequences. When decoding, we predict the output sequence, which will get the following maximum scores, and the sequence with the maximum score will be the output of CRF structure.[11][12]

According to the named entity recognition convention, words are divided into four categories: B-PER, I-PER, E-PER and 0. B-PER indicates that the character is located at the beginning of the entity character boundary, I-PER indicates that the character is located in the middle of the entity character boundary, E-PER means that the character is located at the end, and O indicates that the character is irrelevant to the entity.

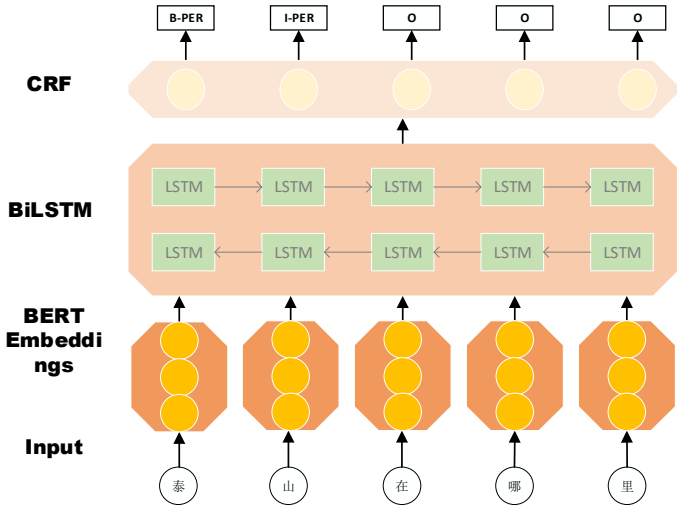


Figure 2. NER with Bert-BiLSTM-CRF

3.3. Query Classification

Relationship extraction is an important part of the question answering system to assist the system to understand the query sentence. In this paper, naive Bayes classification algorithm (NB algorithm) is used to realize simple problem and template matching process. Naive Bayes classification algorithm is based on conditional probability, Bayes theorem and independence hypothesis. The detailed process is as follows: the expression of Bayes' theorem is as follows, where x and Y denote characteristic variables, c_i is the classification label, $p(c_i \mid x, y)$ indicates that the possibility of classification into category c_i when the feature is x, y:

$$p(c_i \mid x, y) = \frac{p(x, y \mid c_i) p(c_i)}{p(x, y)} \tag{7}$$

3.4. Cypher Statement Generator

The main function of this model is to construct cypher query statements to query the answers which only needs to determine the query entity E and the entity relationship r to construct the query[13]: "$match(n:e) - [re:R] \rightarrow (s)$ *return* s", where $N:e$ means to assign the entity name e to the entity n, and

$re:R$ means to assign the relationship name r to the relationship link re. The query statement will query the entity with relation R to entity E and return.[14][15] For example: “Do you know where Mount Tai is?”. To confirm that the entity is "MountTai" and the relationship is "address", it can be “ $match(n:MountTai)-[re:address]\rightarrow(s)\ return\ s$ ”, which will match all the entities has the relationship "address" with “MountTai” and return.

4. Experiments

4.1. Data description

In this paper, after collecting data from the popular tourism websites, such as Horse beehive travel and Baidu travel and through the analysis and preprocessing of massive data. NER based on tourism domain is constructed for this study. The dataset information is shown in Table 1 below.

Table 1 NER dataset description

Dataset Name	Scale
NER based on tourism domain	**32786**
Annotation training data	**21859**
Annotation verification data	**5463**
Annotation testing data	**5464**

Due to the relatively small scale of the data set, in order to make full use of the sample data, this experiment divides the data set into training set, verification set and testing set according to the ratio of 4:1:1, and uses the method of 4-fold cross validation to carry out the experiment. In order to verify the feasibility and significance of the proposed method, a model comparison experiment is also been carried out.

4.2. Environment and parameter setting

The GPU of this model is GTX 2060ti, the operating system is windows 10, the programming environment is Python 3.5, and the framework is tensorflow 1.14.0. In order to ensure the consistency of entity recognition algorithm, the initialization parameters are set: the pre training word vector is 300, the number of GNN neurons is 200, and the initial learning rate is 0.05.

4.3. Evaluation criterion

In this paper, we use precision, recall and F1-score to evaluate the model.

$$P(precision)=\frac{TP}{TP+FP}\times 100\% \tag{8}$$

$$R(recall)=\frac{TP}{TP+FN}\times 100\% \tag{9}$$

Occasionally, if the recall and the precision are both higher, it means the whole performance is better. However, sometimes there is a contradiction between P and R, which means the precision is high but recall is low. As a result, it is proper to use harmonic mean of both P and R. Specially, use F-score instead.[16]

$$F1(fscore)=\frac{2PR}{p+R} \tag{10}$$

4.4. Analysis of experimental results

This paper tests the accuracy, recall and F1-value of the NER model. In order to prove the effectiveness of this model, the existing machine learning methods applied to NER tasks several mainstream deep learning models are used for comparative analysis for the unified testing standard:

(1) CRF: using superposition conditional random field method to solve the problem of tourism information entity nesting.

(2) BiLSTM + CRF: one of the classic models of NER task, general domain recognition effect is very good.

(3) BiLSTM+CRF(+bigram): intend to test the performance of the bigram.

(4) Bert + CRF: the introduction of pre training model, in the field of natural language gradually known as the mainstream model.

(5) Transformer + CRF: transformer has powerful feature extraction ability, and uses transformer to replace RNN model.

(6) HMM: based on the HMM algorithm.

(7) Bert-BiLSTM + CRF: traditional model here used for tourism field.

Table 2 Main results of resume

Models	P(precision)	R(recall)	F1(F-score)
CRF	**87.71**	**87.93**	**87.20**
BiLSTM+CRF	**83.30**	**79.10**	**80.23**
BiLSTM+CRF+(bigram)	**84.20**	**80.23**	**82.31**
Bert+CRF	**85.94**	**86.61**	**86.27**
Transformer+CRF	**74.91**	**82.91**	**78.71**
HMM	**69.12**	**66.34**	**68.20**
Bert-BiLSTM+CRF	**90.23**	**91.23**	**91.10**

From the table II. HMM model is lower than other deep learning models because HMM depends on the current state and its corresponding observation objects. Also, by comparing transformer + CRF model with BiLSTM + CRF model, it can be seen that transformer is not as good as BiLSTM + CRF because transformer is not suitable for NER task in directivity, relative bit size and sparsity. Although the introduction of absolute bit coding to make up for this lack. However, in the task of named entity recognition, the effect is still not ideal.

Additionally, it can be clearly figure out that the evaluation of Bert-BiLSTM-CRF obtains the state of art in the tourism field, each parameter obtains the best performance. Because Bert uses the encoder of transformer to arise the ability of feature extraction and obtains sufficient context information.

5. Conclusion

This research proposes an intelligent question answering system based on tourism knowledge graph, and the core of which includes: 1) the named entity recognition model based on Bert-BiLSTM-CRF, which solves the task of extracting entities from questions. 2) the template classification model based on NB algorithm matches the question with the existing template to obtain the relationship of entities. Experiments show that the QA system constructed in this study obtains state-of-art performance, and it is suitable for multiple regional travel scenarios.

6. Discussion

Although using templates for question classification has many advantages, such as simplicity, high accuracy, wide application and so on. However, the establishment of template library mostly requires experts to manually build it for individual problems, which costs a lot of human resources, and the later

maintenance still acquires manual work; At the same time, the generated template library is highly dependent on the data itself. During the next steps, semantic parsing or graph embedding for relationship matching and automatic templates generation could be a treasure orientation.

References

[1] Turing A M.Computing machinery and intelligence[J].Mind, 1950, 59 (236):433-460.
[2] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine[J]. Communications of the ACM,1966,9(1):36-45.
[3] Tunstall-Pedoe W.True knowledge:open-domain question answering using structured knowledge and inference[J].AI Magazine,2010, 31(3):80-92.
[4] Cui W, Xiao Y, Wang H, et al. KBQA: learning question answering over QA corpora and knowledge bases[J]. Proceedings of the VLDB Endowment,2017,10(5):565-576.
[5] Abujabal A, Yahya M, Riedewald M, et al. Automated template generation for question answering over knowledge graphs[C]. The Web Conference,2017:1191-1200.
[6] Cocco R, Atzori M, Zaniolo C, et al. Machine learning of SPARQL templates for question answering over LinkedSpending[C]. Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2019:156-161.
[7] Berant J, Chou A K, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]. Conference on Empirical Methods in Natural Language Processing,2013:1533-1544.
[8] Kwiatkowski T, Zettlemoyer L, Goldwater S, et al. Lexical generalization in CCG grammar induction for semantic parsing[C]. Conference on Empirical Methods in Natural Language Processing, 2011:1512-1523.
[9] Artzi Y, Lee K, Zettlemoyer L, et al. Broad-coverage CCG semantic parsing with AMR[C]. Conference on Empirical Methods in Natural Language Processing,2015:1699-1710.
[10] Reddy S, Lapata M, Steedman M, et al. Large-scale semantic parsing without question-answer pairs[J]. Transactions of the Association for Computational Linguistics,2014,2(1):377-392.
[11] Reddy S, Tackstrom O, Collins M, et al. Transforming dependency structures to logical forms for semantic parsing[J]. Transactions of the Association for Computational Linguistics, 2016, 4(1): 127-140.
[12] Yih W, Chang M, He X, et al. Semantic parsing via staged query graph generation: question answering with knowledge base[C]. International Joint Conference on Natural Language Processing,2015:1321-1331.
[13] Bast H, Haussmann E. More accurate question answering on freebase[C]. Conference on Information and Knowledge Management, 2015:1431-1440.
[14] Hao Z, Wu B, Wen W, et al. A subgraph-representation-based method for answering complex questions over knowledge bases[J]. Neural Networks,2019,119:57-65.[1]–[4]
[15] Dongjin H, Han Q, et al. film original sound intelligent question answering system based on bet-cnn, computer technology and development, Vol. 30, No. 11, pp. 158-162, 2020
[16] Siyu W, Jiangtao Qin, et al. "Research on online product Q & A based on knowledge map", Chinese Journal of information, Vol. 34, No. 11, pp. 104 – 112, 2020.