An Intelligent Question Answering System of the Liao Dynasty Based on Knowledge Graph

Shuang Liu¹ · Nannan Tan¹ · Hui Yang² · Niko Lukač³

Received: 8 May 2021 / Accepted: 19 August 2021
© The Author(s) 2021

Abstract

The Liao Dynasty was a minority regime established by the Khitan on the grasslands of northern China. To promote and spread the cultural knowledge of the Liao Dynasty, an intelligent question-and-answer system is constructed based on the knowledge graph in the historical and cultural field of the Liao Dynasty. In the traditional question answering system, the quality of answers was not high due to incomplete data and distinctive vocabulary. To solve this problem, a combination method of Liao Dynasty question-and-answer database and KB is proposed to realize knowledge graph question answering, and a joint model of Siamese LSTM and fusion MatchPyramid is proposed for semantic matching between questions in the question-and-answer database. With the joint model, it is easy to perform semantic matching by fusing sentence-level and word-level interactive features through LSTM and MatchPyramid. Furthermore, the question sentence with the same semantics as the user input question sentence is retrieved in the question-and-answer database, and the answer corresponding to the question sentence is returned as the result. The experimental results show that our proposed method has achieved relatively good performance in the historical domain of the Liao Dynasty and the open-domain knowledge graph, and improved the accuracy of question and answer.

Keywords Knowledge graph · Question answering · Siamese LSTM · MatchPyramid · Deep semantic matching

1 Introduction

With the rapid development of Internet Technology, information resources have gradually increased. For users, the ways to obtain information have become diverse and convenient. Among them, online retrieval through search engines is

an important way to obtain information. However, traditional search engines rely on the amount of web page information on the Internet. Although searching results from the engines can cover the results of user questions, users still need to select the answers they need from the returned web pages. In addition, the data sources of traditional search engines often come from unstructured data such as web documents. With this method, it is difficult to adapt to the trend of rapid information growth. Therefore, the knowledge graph technology came into being.

As an emerging data storage method, the knowledge graph is stored in the resource description framework (RDF) triple format [1]. The concept of knowledge graph was proposed by Google in 2012 [2], which enables the search method to find the knowledge information required by users more accurately, more extensively, and more continuously than other methods. Subsequently, knowledge graph technology developed rapidly in industry and academia, and knowledge graphs in many general fields have been produced one after another, such as English knowledge base YAGO [3], Probase [4], DBpedia [5], Chinese knowledge base zhishi.me [6], Ownthink [7], XLORE [8], etc. The emergence of

✉ Shuang Liu
liushuang@dlnu.edu.cn

Nannan Tan
tannn761@gmail.com

Hui Yang
yanghuimail2020@gmail.com

Niko Lukač
niko.lukac@um.si

¹ School of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China

² School of Hotel Management, Nanjing Institute of Tourism and Hospitality, Nanjing 211100, China

³ Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroska cesta 46, 2000 Maribor, Slovenia

these knowledge graphs makes automatic question and answer based on knowledge graphs feasible.

The goal of the knowledge base question answering (KBQA) [9] is to accept the user's natural language questions, and then find the corresponding one or more triples in the knowledge graph to introduce knowledge after understanding the semantics of the question sentence, and help to complete the answer. However, how to understand the deeper semantics of natural language and how to integrate natural language with structured knowledge have become the most challenging task at present. To promote the history and culture of the Liao Dynasty in China, an intelligent question answering system is constructed for the history and culture of the Liao Dynasty based on the knowledge graph technology. The question answering system allows users to obtain accurately historical information of the Liao Dynasty more conveniently and quickly. Knowledge Graph uses data processing technology to fuse massive amounts of historical unstructured text data of the Liao Dynasty and transform it into semantically rich structured data composed of domain entities, attributes, and their interrelationships. It also provides a high-quality data source for intelligent question answering system to improve the accuracy of data query effectively.

In recent years, many researchers have combined continuous text knowledge and knowledge graphs as a data source for KBQA. For example, Han et al. [10] proposed a question-and-answer model based on text-enhanced knowledge graphs, which encoded entities in the KB through text information and applied graph convolutional networks [11] to perform reasoning on KB. Das et al. [12] proposed a general model that used a combination of text and KB, combining with a memory network to complete question and answer. Xiong et al. [13] assisted entity coding with the help of graph attention mechanism to enhance the understanding of questions and documents so as to obtain answers.

In fact, the content of the KB is usually incomplete and imperfect, and it is not enough to cover all the knowledge that users need. There are a large number of unstructured texts on the Internet that can easily cover knowledge that does not exist in the KB. Therefore, many scholars choose to apply the method of combining unstructured text and KB in question-and-answer research. However, there are great challenges in applying this combination form to automatic question answering systems. First, how to express the meaning of question sentences when the structure of the KB and the text are inconsistent, which has greater limitations in text corpus recognition. Second, the performance is significant in ensuring a single text paragraph containing the answer. However, when there are multiple large paragraphs of text and whether the answer appears in the text is unknown, the amount of calculation increases and the desired result cannot be obtained, resulting in the performance of the model

Decline. In response to the above problems, external text knowledge is introduced to construct a question-answer database (QAB) containing question-and-answer pairs to replace the huge unstructured text data. Then, a joint model of Siamese LSTM and fusion MatchPyramid is proposed for deep text matching to realize KBQA. Here, an improved fusion MatchPyramid model is adopted, which fully considers the context information of the words in the question sentence and the importance of the core words. In the model, the text which is semantically matched with the question input by the user is the question in the QAB. Specifically, the method proposed in this paper uses the Siamese LSTM and the fusion MatchPyramid model to extract different question feature combinations and then judges whether the semantics of the two questions is the same, and then retrieves the answer corresponding to the question from the QAB. Question features include the contextual information of a single question and the interactive information between two sentences. Extensive experiments are carried out based on the method of information retrieval using the QAB and the automatic learning model of question sentences. The experimental results show that our method is better than other baseline models in the data set of the historical and cultural fields of the Liao Dynasty.

The main contributions of this paper are as follows: (1) A hybrid model is proposed to solve the problem of text semantic similarity, which improves the performance on the basis of the MatchPyramid model. (2) A QAB is constructed to improve the accuracy of automatic question and answer through a hybrid data source combining the knowledge base and the QAB. (3) A knowledge graph is constructed, and NER dataset and question answer dataset in the historical and cultural field of the Liao Dynasty are also completed.

The rest of this paper is arranged as follows. In Sect. 2, we introduce the related work of KBQA and text semantic matching. In Sect. 3, we introduce the background of the proposed method and the overall implementation of QA. In Sect. 4, we introduce the realization of QA system in the historical field of Liao Dynasty. In Sect. 5, we give specific experiments and analyses. In Sect. 6, we make a summary of this paper.

2 Related Work

At present, the most mainstream methods in KBQA at home and abroad can mainly be categorized into two types, which are the method based on semantic analysis and the method based on information retrieval. The method based on semantic analysis first parses the natural language question sentence into a series of structured logical representations. Then, it parses the generated series of logical forms from the bottom-up to generate a logical form expressing

the semantics of the entire sentence, which can be directly obtained by querying in the KB answer. Yuk et al. [14] relied on artificially labeled logical expressions as supervision information. The limitation of using this method is that the labeled data only include a small number of relational predicates. Many researchers have used new methods that do not require manual creation of templates to solve this problem [15–17]. Nevertheless, this method emphasizes the use of logical forms in the implementation process. Although it reflects the explanatory and reasoning nature of the question-and-answer process to a certain extent, the pros and cons of this type of method are too dependent on the generation of logical forms and quality at analytical time, which in turn causes information loss.

For the second information retrieval method, the subject entity of the natural language question is first determined, and several candidate triples of the question are generated, and then, the candidate triples are scored and sorted using the deep learning model. Finally, the triple information that best matches the question is obtained as the predicted result. The overall framework of this type of method is simple with good generalization capabilities, not requiring rules or features designed by men. Therefore, many researchers presented improvements on the basis of this method. Yu et al. [18] proposed a method to enhance relationship matching, using two-layer BiLSTM to match candidate relationships at multiple levels. Zhu et al. [19] proposed the Tree2Seq method, which used Tree-based LSTM to encode the context information of entities and relationships in Query Graph. In the decoding process, a mixture of generation mode and reference mode is used to capture the level of association between the query and the question. Dong et al. [20] proposed a MCCNNs question answering model using the Freebase base. This method automatically analyzes questions from three different aspects, which improved KBQA's effect by considering the word order information and the relationship between the question and the answer. Chen et al. [21] proposed one BAMnet model, which achieved good results on the WebQuestion dataset [24] by relying on the Attention mechanism [22] and MemNN [23] without relying on manual construction of templates.

Text semantic matching is a basic technology in natural language processing [25]. It aims at analyzing and determining the degree of semantic relevance between two texts. It is widely used in natural language reasoning, question answering, semantic understanding, and many other fields. At present, the mainstream model of text semantic matching is the deep text matching model, which is relatively more competitive on the corresponding data set. For example, the object in the DSSM [26] model is the similarity between the query content and the document. Each document object and query content are vectorized and coded through a five-layer neural network, and then, the output results are spliced and

input into one fully connected layer to compute the similarity between the query content and the document. The BiMPM model proposed by Wang et al. [27] simultaneously uses four matching strategies to perform semantic similarity matching at multiple angles, which solves the problem of insufficient interactive matching. The ESIM model proposed by Chen et al. [28] effectively combines BiLSTM [29] and Attention, and further realizes global inference based on local inference, and obtains the semantic relationship of the text through inference.

Different from the existing work, one combining method is presented based on information retrieval and the deep text matching model in automatic question answering. A joint model of Siamese LSTM-fusion MatchPyramid is proposed, which is used to fuse two different levels of feature information in the feature extraction part. In this paper, Siamese LSTM [30] is used to generate the context vector of the question entered by the user and the question in the QAB. This paper proposes an improved fusion MatchPyramid model, which fully considers the context information of each word-level matching and the importance of the core words on the basis of the single-channel MatchPyramid model [31], and calculates the semantic information of the word interaction between two different questions. In this way, the rich information in the structured KB and the unstructured QAB can be applied to the QA system.

3 Background

The realization of intelligent question answering based on the knowledge graph of Liao Dynasty history and culture is mainly divided into the following two modules. The first module is the building module of the knowledge graph and the second module is the realization module of automatic question answering. Here, the building steps of the knowledge graph include knowledge extraction, knowledge fusion, and knowledge storage. The focus of this article is on the automatic question-and-answer module. This module combines deep text matching methods on the basis of traditional information-based retrieval methods. Aiming at the information loss of a single model, a multi-channel fusion MatchPyramid model is proposed based on the currently widely used algorithm. This model can obtain the matching scores and contextual information of the words, respectively. Then, the Siamese LSTM model and the fusion MatchPyramid model are spliced together to fully obtain sentence-level and word-level information. This idea can use external data to solve the problem of incomplete storage of triples in the KB and inaccurate identification based on information retrieval methods. Among them, the named entity recognition part of the information retrieval method uses the BiLSTM-CRF (Bi-directional Long Short-Term

Fig. 1 Automatic question answering module overall flow chart

Memory networks-Conditional Random Field) [32] model, the question intention recognition uses the TextCNN [33] model, and the deep text matching part uses the Siamese LSTM-fusion MatchPyramid model. The overall process of the automatic question answering module is shown in Fig. 1.

3.1 Construction of Knowledge Graph in Liao Dynasty

After understanding and analyzing the history and culture of the Liao Dynasty, crawler technology is used to obtain original corpus from Baidu Encyclopedia and historical websites, and entity extraction tools and relationship classification models are used to automatically extract entities and relationships from semi-structured and unstructured data. Here, the entity types consist of Chinese name, nickname, capital, historical figure, language, ethnicity, military deployment, enforcement system, art form, technology, diplomatic contacts, clothing and hair accessories, business contacts, population, religion, folklore, architecture, territory division, food, festivals, funeral taboos, etc. Each entity category contains multiple entities. After knowledge fusion and data storage, a historical and cultural knowledge graph of the Liao Dynasty containing more than 15,000 entities was constructed. The construction process of the Liao Dynasty historical knowledge graph is shown in Fig. 2.

3.2 QAB Construction

Because the Liao Dynasty was founded by Chinese ethnic minorities, its history and culture contain strong national cultural characteristics. Due to the impact of modern civilization, many ancient civilizations have begun to decline. As a result, some words and information in the Liao Dynasty became unknown. This led to the fact that some vocabulary and information in the Liao culture became gradually unknown, which made the storage of the Liao

Fig. 2 The flowchart of historical knowledge graph construction in Liao Dynasty

Dynasty KB insufficient and poor storage quality. Because, the data volume and quality of the KB play a vital role in the realization of question and answer. Therefore, based on the above analysis, a QAB of the Liao History is constructed, and the acquired external knowledge is used to further make up for the low accuracy of question and answer caused by the lack of entities in the KB. And based on the traditional question-and-answer system framework for information retrieval, a deep text matching model is proposed. The model can use the combined knowledge source of the KB and the QA database to make up for the problem of unclear question semantic recognition and to supplement the lack of data in the KB, so as to solve the problem that the QA is limited by the capacity of the KB.

In the construction of the QAB, crawler technology is adopted to crawl the question, the answer, answer time, and number of likes of the corresponding question-and-answer page in professional question and answer websites, historical websites, forums, and post bars. Then, the crawled data are preprocessed and filtered according to the corresponding rules formulated. Based on people's reading behavior, the users usually give likes for comments that they think are organized and well founded when they browse the pages. Therefore, according to the answer time and the number of likes in the question and answer website for one-question, multiple-answer questions, the top three answers are selected in the number of likes (if the number of likes is the same, the earlier answer is considered) and stored it in the QAB. Then, the title of the question is marked according to the subject of the question. The final QAB storage format can be (titles, question, answers, number of likes).

4 Model Description

In the KBQA implementation process, the named entities and relationships are first obtained through named entity recognition and question intention recognition, and then, the cypher sentence is used to query in the Neo4j graph database [34], and the corresponding entity or attribute value is found and returned as the answer. If the corresponding entity or attribute information is not retrieved in the KB, then deep text matching is performed.

A deep text matching model for KBQA is proposed here. The basic idea is as follows. First, the question sentence and the representation vector of the question sentence in the QAB are initialized. Then, they are sent, respectively, to the Siamese LSTM and the fusion MatchPyramid model to obtain the semantics between sentences information. In addition, the MatchPyramid model is improved by merging multiple channels, and a fusion MatchPyramid model is proposed, which combines two channels containing the importance of words and context information, and then extracts high-level information through the convolutional layer and the pooling layer. Finally, the feature information extracted from the two models is spliced and classified. This method makes full use of the information of external text data, improves the use of KB information, and effectively improves the quality of question and answer. The specific introduction of the model is as follows.

4.1 Siamese LSTM-Fusion MatchPyramid Model

A deep text matching method combining Siamese LSTM and the fusion MatchPyramid model is proposed here. The structure diagram of the model is shown in Fig. 3. Specifically, the two models are combined to extract the semantic features of the question from two different perspectives of the sentence and the word, and then, these features are combined and classified to obtain the answer to the question. Next, the working process of the model is explained in detail.

4.1.1 Siamese LSTM

The Siamese neural network [35] is a feature similarity measurement method proposed by Yann Lecun, which is widely used in the analysis of similar samples. The Siamese neural network is composed of two neural networks with similar or identical structures. The two networks share weights and thresholds, and they are trained separately. Then, the outputs of the two networks are spliced into a fully connected layer. With the Siamese network, the parameters that need to be trained can be reduced while effectively preventing the occurrence

Fig. 3 Siamese LSTM-fusion MatchPyramid model diagram

of over-fitting. LSTM solves the problems of gradient disappearance and gradient explosion prone to occur in RNN [36] by improving the network structure of RNN. Thus, LSTM can remember long-term historical information and improve the long-term dependence of RNN. LSTM has been widely used in various fields of research, and has achieved good results. Therefore, LSTM is used as the two sub-networks of the Siamese neural network.

1. Feature input layer. First, the input text vector of the user input question sentence and the question sentence in the QAB are obtained, respectively, through the pre-training word vector. Here, GloVe(Global Vectors) is used to initialize each vector [37]. The question entered by the user is defined as $q = (v_1, v_2, \dots, v_n)$, and the question in the QAB is as $Q = (w_1, w_2, \dots, w_m)$, where v_i and w_j represent the word vectors of the i and j words of the $h_m^{(b)}$ question, respectively. Then, the word vectors are sent to the two LSTM networks respectively.
2. Extraction layer. The LSTM layer uses the obtained word vectors for encoding to obtain fixed-length feature vectors for similarity analysis. The LSTM layer contains multiple LSTM units, and the input of each calculation unit is the value x_t^* at a certain time in the sequence, the hidden state h_{t-1} , and the memory unit C_{t-1} output by the calculation unit at the previous time. Here, the standard LSTM model is adopted, and its calculation process is shown in the following formula:

$$f_t = \sigma(W_f \cdot [h_{t-1}, X] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, X] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, X] + b_C) \quad (3)$$

$$C_t = f_t^* C_{t-1} + i_t^* \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, X] + b_o) \quad (5)$$

$$h_t = o_t^* \tanh(C_t). \quad (6)$$

3. Output layer. Here, the output layer does not adopt the similarity measurement method of the traditional siamese neural network model. Instead, the last hidden state of the LSTM layer is directly used as the output.

4.1.2 Fusion MatchPyramid

The above-mentioned Siamese LSTM network extracts coding vectors from sentences at the sentence level. However, in deep text matching, it is necessary to obtain as much semantic information as possible to judge the similarity between sentences more accurately. Taking text matching as an example, two questions are given with the same semantics as follows.

S1: Did the Khitan in the Liao Dynasty only have these two surnames?

S2: Are Yelu and Xiao the only two surnames in the Liao Dynasty?

For the above two sentences, in the Siamese LSTM, it will judge whether they are similar according to the overall meaning of the two sentences. However, when the user does not fully understand the semantics of the text, it will lead to a certain degree of misjudgment. Therefore, this paper considers adding the MatchPyramid model to increase word-level semantic information. For example, “Liao” in S1 and “Liao” in S2, “surname” in S1 and “surname” in S2, etc., belong to word-level matching, and “Liao Dynasty” in S1 and “Liao Dynasty” in S2. The “two surnames” in S1 and the “two surnames” in S2 are phrase-level matching. Finally, sentence-level matching information can be obtained based on the same words and similar meanings at the phrase level to determine whether the meanings of the two questions are the same.

The MatchPyramid model can effectively focus on word-level relationships by constructing a similarity matrix. However, it still has shortcomings. First, only the matching score is considered, ignoring the importance of core words. Second, the contextual information is ignored at the word level.

Based on the above analysis, MatchPyramid model is improved here by integrating the importance of words and context information into the MatchPyramid model to complete text matching. Then, the convolutional layer and the

pooling layer are accessed to obtain the feature vector. In this paper, BiLSTM is used to extract the context information of words, and the important information of core words is realized by a layer of feedforward neural network. The overall structure of the model is shown in Fig. 4.

To model text matching as image recognition, the key is to solve the problem of text and image representation. To solve this problem, the text vector obtained by GloVe is used to perform a pairwise interactive dot product operation on the words v_i and w_j of the two question sentences to obtain word-level matching information to form a similarity matrix. The formula is as follows:

$$M_{i,j} = v_i^T w_j. \quad (7)$$

It represents the similarity matrix of the two questions.

In terms to the importance of core words, the feedforward neural network method is used to obtain the formula as follows:

$$M_i^{(a)} = \sigma(wx_i + b) \quad (8)$$

$$M_i^{(b)} = \sigma(wy_i + b). \quad (9)$$

Here, $M_i^{(a)}$ represents the importance of the i th word of sentence a , and w and b are the feedforward neural network parameters initialized randomly. During the training process, M will be continuously updated to learn the importance of words. It can be seen from the formula that the importance at this time is only related to the word itself, and the same word will get the same score in different contexts. However, the importance of words and matching scores should also consider the context.

Fig. 4 Fusion MatchPyramid model diagram

To consider the context information of words, BiLSTM is adopted to encode the text, and then, the hidden layer vector output is used at each time step to replace the word vector for subsequent interactive dot product operations. The coding process formula of the two questions is as follows:

$$x_1, x_2, \dots, x_m = \text{BiLSTM}(x, \theta) \quad (10)$$

$$y_1, y_2, \dots, y_m = \text{BiLSTM}(y, \theta). \quad (11)$$

The output hidden layer vector has the same dimension as the text vector, so the dot product operation is performed on the encoded hidden layer vector to obtain the similarity matrix. The formula is as follows:

$$M'_{ij} = x_i^T y_j \quad (12)$$

The word importance formula obtained through the encoded hidden layer vector is as follows:

$$M'_i = \sigma(wx_i + b) \quad (13)$$

$$M'_j = \sigma(wy_j + b). \quad (14)$$

Then, the similarity matrix and context similarity matrix are concatenated to obtain a fusion similarity matrix M_{fus} , and the formula is as follows:

$$M_{fus} = M_{ij} \oplus M_i^{(a)} \oplus M_i^{(b)} \oplus M'_{ij} \oplus M'_i \oplus M'_j, \quad (15)$$

Where $\oplus$ represents the splicing between vectors. This model uses a convolutional neural network to extract matching information at different levels. The convolution is calculated as follows:

$$M_{ij}^{(l,k)} = \sigma \left(\sum_{s=0}^{k-1} \sum_{t=0}^{N-1} w_{s,t} \cdot M_{fus+s,j+t} + b \right), \quad (16)$$

where w is the weight matrix, b is the bias, and $M_{ij}^{(1,k)}$ is the output matrix obtained by the k th convolution kernel, which is the activation function.

4.1.3 Joint Model

Because the Siamese LSTM only considers the context information of a single sentence and ignores the connection between two sentences, the fusion MatchPyramid pays more attention to extracting the interactive information between the two sentences. Therefore, these two models are combined to consider the semantic similarity of question sentences in the context of a single sentence and the interactive information between sentences. The specific introduction of the model is as follows:

1. Embedding layer: using GloVe for initialization.
2. Feature extraction layer: the high-order features extracted from the fusion MatchPyramid model and the context features generated by the Siamese LSTM are spliced.
3. Output layer: inputting the combined features to the fully connected layer for classification. The Softmax function is used to predict the probability of each tag category. In the Siamese LSTM-fusion MatchPyramid model, the loss function is set as the cross-entropy loss function. The cross-entropy loss function can enable the learning rate of the model to be controlled by the output error, and the convergence speed is fast, which can effectively avoid the problem of low learning rate. At the same time, the model does not fall into a local optimal solution. The formula for cross entropy is as follows:

$$L_i = - \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log (1 - p_i)], \quad (17)$$

where y_i represents the real label category, and p_i represents the predicted label category.

5 Experiment

To evaluate the performance of the question answering system, the models are tested in the historical and cultural fields of the Liao Dynasty and the open domain. Accuracy (Acc) is adopted as the evaluation criteria of algorithm performance. It is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}. \quad (18)$$

5.1 Experimental Setup

5.1.1 Datasets

An intelligent question answering system for the historical and cultural fields of the Liao Dynasty is constructed in this paper, and a historical and cultural knowledge graph of the Liao Dynasty containing more than 10,000 entities is constructed based on structured, semi-structured, and unstructured data. In the process of constructing the question answering system, the following data sets are mainly used for model training and testing:

1. Liao Dynasty Question-and-Answer Dataset: The experimental dataset in this article is composed of 10,000 questions related to the history and culture of the Liao Dynasty, which are artificially constructed and labeled.

- Liao Dynasty History QAB: More than 70,000 answers to questions related to the history of the Liao Dynasty were crawled from historical QA website (<https://www.allhistory.com/>), Sina Aiwen (<https://iask.sina.com.cn/>), and other website, which are used for semantic matching between questions.
- WebQA Dataset [38]: In the dataset, each sample is composed of original text, questions, and answers, among which the questions are raised based on the original text. There are 36145 items in the training set, 3024 items in the test set, and 3018 items in the verification set.
- CN-DBPedia [39]: This KB is a large-scale general domain structured encyclopedia developed and maintained by the Fudan University Knowledge Workshop Laboratory. It contains 9 million encyclopedia entities and 67 million triple relationships.
- baike2018qa [40]: It contains 1.5 million pre-filtered, high-quality questions and answers. In this article, only the questions and answers in the dataset are selected and stored in the QAB.

5.1.2 Parameter Setting

The model parameters used in the article are set as follows, the batch size is set to 64, the learning rate 0.001, the drop-out 0.5, the number of filters 2, and the Adagrad optimization method is used for optimization. In the word vector part, a 200-dimensional word embedding trained by the GloVe model is used. The dimensionality of the hidden presentation layer in the Siamese LSTM is set to 200. In the fusion MatchPyramid model, the size of the convolution kernel is set to 3, and the number of hidden layer neurons in BiLSTM is 100.

5.2 Results and Analysis

5.2.1 Siamese LSTM-Fusion MatchPyramid Model

To prove the effectiveness of the proposed model in KBQA, the constructed Liao Dynasty historical knowledge graph and the Liao Dynasty question-and-answer dataset are used in the following models to complete the comparative experiment. The experimental results are shown in Fig. 5.

- Zhang et al. [41] constructed a fine-grained knowledge base question-and-answer model based on BiLSTM-CRF and N-Gram algorithms. It uses the attention mechanism and CNN to capture the relationship between the main entity of the question and the candidate relationship, and the relationship between the question and the original words of the relationship from the semantic level and the word level, respectively.

Comparison of all experiment result(Acc%)

Fig. 5 Comparison results with other models

Comparison of all experiment result(Acc%)

Fig. 6 Comparison of our method and the baseline model in the open domain

- Luo et al. [42] proposed a relationship detection method using a multi-angle attention mechanism, which uses the attention mechanism to extract different levels of semantic relations of question patterns and relationships from different angles.
- Wang et al. [43] proposed a Siamese Attention Network (Siamese Attention Network) to extract the characteristics of problems and attributes. Rule reasoning is also introduced to improve KBQA's answering ability.

In the above experiments, the first three comparison models did not consider the question of using the QAB for semantic matching, so a comparative experiment is also added that does not use the QAB for semantic matching, namely Ours (Unused deep semantic matching). In addition, a comparison experiment using the QAB for semantic matching is also added. The comparison model is different from our model in that it does not use multi-channel fusion to obtain information. To facilitate the distinction, here, we name the comparison model Ours-1 (Siamese LSTM-MatchPyramid), and our model is named Ours-2. From the model comparison histogram in Fig. 6, it can be clearly seen that when only the KB

is used to implement question and answer, the performance of the Ours (Unused deep semantic matching) method in this article is lower than the other three models, because the QAB is not used to make up for the lack of KB and the performance degradation caused by model recognition errors. The main errors in the experiments of the first four models are: the inability to retrieve information through multi-hop reasoning for complex questions, difficulty in identifying historical professional terms in the entity recognition module, and insufficient data in the KB and QAB. All these factors affected the performance of QA. However, when the QAB for deep text matching is introduced to make up for the shortcomings of the QA model, the proposed model can achieve good results with an accuracy rate of 81.63%, which is 2.48% higher than the result of using the Siamese LSTM-MatchPyramid model for QA. After analysis, it is found that when the storage of the KB is insufficient, the problem of insufficient knowledge coverage can be made up with the help of external knowledge. In addition, the Siamese LSTM-fusion MatchPyramid method proposed in this paper can fully obtain word-level information and semantic-level information, and has strong learning ability. It can solve the information loss caused by MatchPyramid single-channel matching information, thereby improving the performance of question answering.

To verify the generalization of the open-domain question answering, CN-DBPedia [38] knowledge base and WebQA [39] (question, answer) are used to validate the above comparison model. Among them, the QAB for deep text matching with question sentences uses the baike2018qa dataset [40]. The result is shown in Fig. 6. Through experiments, it is found that our model is 5.8% higher than the method of Zhang et al., about 4% higher than the method of Luo et al., 2.74% higher than the model of Wang et al., and 6.44% higher than Ours (Unused deep semantic matching) method. After analysis, it is found that the above-mentioned comparison model can achieve ideal results in simple single-hop problems, but in the face of complex multi-hop problems, the relationship cannot be accurately derived, so the effect is not good. However, the model proposed in this paper can also match the question sentence with the question sentence in the QAB in complex questions, so as to retrieve the correct answer. Therefore, the model in this paper can also achieve better results in open-domain question answering.

To verify the performance of the model in the open-domain question answering, the KB is not sufficient. This article randomly deletes the triples of CN-DBPedia knowledge base with a probability of 50% to form a new KB. The result is shown in Table 1. The results on WebQA show that when the KB is insufficient, the accuracy rates are, respectively, 11.66%, 10.78%, and 8.82% higher than the baseline model. Therefore, the method in this paper

Table 1 Comparative experiment when the KB is insufficient in the open-domain question answering

Method	50%KB+QAB	100%KB+QAB
Zhang et al.	57.62	79.47
Luo et al.	58.5	81.26
Wang et al.	60.46	82.53
Ours-2	69.28	85.27

Table 2 Experimental results of semantic matching

Method	Acc%
DSSM	78.48
ESIM	81.39
Siamese LSTM	82.60
Siamese CNN+MatchPyramid	84.71
Siamese LSTM-MatchPyramid	85.82
Ours-2	87.45

improves the prediction effect. The main reason for the performance degradation of the comparison model is that the KB is insufficient. Many answers need to be obtained through links with more than three sides, and these methods for simple questions cannot find the correct answer. This method can prove that even if there are no multiple edges connected, the correct answer can be found through the QAB.

5.2.2 Deep Text Matching Model Comparison Experiment

To verify the effectiveness of our proposed Siamese LSTM-fusion MatchPyramid method in semantic matching, comparative experiments with the DSSM model, ESIM model, and Siamese CNN-MatchPyramid as the baseline model are also conducted on the Liao Dynasty historical QAB dataset. It can be seen from Table 2 that our model is better than the baseline model, and outperforms the best baseline model with a ratio of 2–3%. After analysis, it is found that single semantic models such as DSSM and LSTM can only obtain one type of feature of the text, which delays the interaction between the two questions. Only using MatchPyramid will result in information loss, while using fusion MatchPyramid to fully consider word importance information and contextual semantic information can effectively improve the problem of information loss. Therefore, the two models are combined together in this paper. Using this model, contextual features extracted through LSTM are merged with MatchPyramid to obtain the interactive information of the question, which can enrich the features extracted by the model and obtain better matching results.

Fig. 7 Experimental comparison results under different QAB settings

5.2.3 Ablation Experiment

To study the influence of each part of the Siamese LSTM-fusion MatchPyramid model on the results of KBQA, we conducted the following experiments. Random sampling of 50%, 70%, and 100% of the QAB is performed to determine the impact of the QAB on the answers when the KB is complete. Figure 7 lists the experimental results of each method. It can be seen from the experimental results that when the knowledge base is complete, the QAB plays a vital role in the retrieval of answers. This method can find the corresponding answer in the QAB by matching the semantic relationship between the question sentences. Through this experiment, it can be found that the performance of the model used in this article is about 0.6%–3% higher than the five mechanisms in the comparison experiment in the case of KB+50%QAB. In the case of KB+100%QAB, it is improved by 1%–4%. In KB+100% QAB, an increase of 3%–6%. It can be found that the fusion proposed MatchPyramid model can obtain more fine-grained word-level information in the question. Combining the Siamese LSTM and the fusion MatchPyramid model can fully consider the sentence-level and word-level features to compensate for the two models to improve the accuracy of QA.

The above experiments are also conducted in the open-domain question answering, and the experimental results are shown in Fig. 8. From experiments results, it can be found that the model in this paper has a good performance in the KBQA base on the basis of constructing the QAB with the help of external knowledge.

6 Conclusion

In this paper, a heterogeneous data source is used for combining with a knowledge base and a QAB to implement knowledge graph question answering. A deep semantic

Fig. 8 Experimental comparison results under different QAB settings in the open field

matching model based on Siamese LSTM and fusion MatchPyramid is proposed to assist in the realization of KGQA in the field of the Liao Dynasty. It is introduced that the model learns different feature representations of questions from the Siamese LSTM network and the fusion MatchPyramid model, and integrates the important information of the words and context information for multi-semantic expression. After a large number of experiments, it is shown that the model in this paper can effectively improve the performance of the question answering system in the Liao Dynasty historical field dataset and the open-domain dataset.

Acknowledgements This research was funded by Social Research Project on Economic Development in Liaoning Province (Grant No. 2021slslybkt-022), Natural Science Foundation of Liaoning Province, China (Grant Nos. 20180550921 and 2019-ZD-0175), and Scientific Research Fund Project of the Education Department of Liaoning Province (LJYT201906).

Declarations

Conflict of interest The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ji, S., Pan, S., Cambria, et al. S.E.: A survey on knowledge graphs: representation, acquisition and applications. [arXiv:2002.00388](https://arxiv.org/abs/2002.00388)
- Amit, S.: Introducing the Knowledge Graph: Things, Not Strings, Official Blog (of Google) (2012)
- F.M. Suchanek, G. Kasneci, G. Weikum, YAGO: A large ontology from wikipedia and wordNet[J], *Journal of Web Semantics*, 6 (2008), 203–217
- W. Wu, H. Li, H. Wang, et al., Probase: A probabilistic taxonomy for text understanding. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data (ICMI)*, Scottsdale, USA, 2012, pp. 481–492
- C. Bizer, J. Lehmann, G. Kobilarov, et al., DBpedia-A crystallization point for the web of data[J], *Journal of Web Semantics*, 7 (2009), 154–165
- Niu, X., Sun, X., Wang, et al., H.: Zhishi. me-weaving chinese linking open data. In: *International Semantic Web Conference (ISWC)*, Bonn, Germany, pp. 205–220 (2011)
- MrYener, OwnThink Knowledge Graph. Available online: <https://www.ownthink.com/>
- Wang, Z., Li, J., Wang, Z., et al.: XLORE: A large-scale english-chinese bilingual knowledge graph[C], *International Semantic Web Conference (ISWC)*, pp. 121–124. Australia, Sydney (2013)
- Lan, Y., Jiang, J.: Query graph generation for answering multi-hop complex questions from knowledge bases. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, pp. 969–974 (2020)
- J. Han, B. Cheng, X. Wang, Open domain question answering based on text enhanced knowledge graph with hyperedge infusion[C], *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1475–1481
- Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: *Advances in Neural Information Processing Systems 29 (NIPS)*, Barcelona, Spain, pp. 3837–3845 (2016)
- Das, R., Zaheer, M., Reddy, S., et al.: Question answering on knowledge bases and text using universal schema and memory networks[C], In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 358–365. Canada, Vancouver (2017)
- Xiong, W., Yu, M., Chang, S., et al.: Improving question answering over incomplete kbs with knowledgeaware reader[C], In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4258–4264. Italy, Florence (2019)
- Y.W. Wong, R. Mooney, Learning synchronous grammars for semantic parsing with lambda calculus[C], In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007, pp. 960–967
- Abujabal, A., Yahya, M., Riedewald, M., et al.: Automated template generation for question answering over knowledge graphs[C], In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pp. 1191–1200. Australia, Perth (2017)
- Z. F. Hao, B. Wu, W. Wen, et al., A subgraph-representation based method for answering complex questions over knowledge bases[J], *Neural Networks*, 119(2019), 57–65
- Li, D., Mirella, L.: Coarse-to- fine decoding for neural semantic parsing[C], In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 731–742. Australia, Melbourne (2018)
- Yu, M., Yin, W.P., Hasan, K.S., et al.: Improved neural relation detection for knowledge base question answering[C], In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 571–581. Canada, Vancouver (2017)
- S. Zhu, X. Cheng, S. Su, Knowledge-based question answering by tree-to-sequence learning[J], *Neuro computing*, 372 (2020), 64–72
- Li, D., Furu, W., Zhou, M., et al.: Question answering over freebase with multi-column convolutional neural networks[C], *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 260–269. China, Beijing (2015)
- Chen, Y., Wu, L.F., Zak, M.J.: Bidirectional attentive memory networks for question answering over knowledge bases[C], *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota (NAACL), Minneapolis, Minnesota, pp. 2913–2923 (2019)
- Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate[C], *3rd International Conference on Learning Representations*, pp. 1–15. San Diego, USA (2015)
- Weston, J., Chopra, S., Bordes, A.: Memory networks. [arXiv:1410.3916](https://arxiv.org/abs/1410.3916) (2014)
- Berant, J., Chou, A., Frostig, et al., R.: Semantic parsing on freebase from question-answer pairs. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, USA, pp. 1533–1544 (2013)
- Weng, Z. Q., Zhang, L.: Text semantic matching model based on multi-angle information interaction. *Comput Eng.* <https://kns.cnki.net/kcms/detail/31.1289.TP.20201103.1813.007.html>
- Huang, P., He, X., Gao, J., et al.: Learning deep structured semantic models for web search using clickthrough data[C], *Proceedings of the 22nd ACM International Conference on Conference on Information (ICMI)*, pp. 2333–2338. Australia, Sydney (2013)
- Wang, Z., Hamaza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences[C], *Proceedings of the 26th ACM International Conference on Conference on Information (ICMI)*, pp. 4144–4150. Scotland, Glasgow (2017)
- Chen, Q., Zhu, X.D., Ling, Z.H., et al.: Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1657–1668. Canada, Vancouver (2017)
- Lample, G., Ballesteros, M., Subramanian, et al., S.: Neural architectures for named entity recognition[C], *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California, 2016, pp. 260–270
- Lv, C., Wang, F. P., Wang, et al., J. H.: Siamese multiplicative LSTM for semantic text similarity[C], *ACAI 2020: 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*, Sanya, pp. 1–5 (2020)
- Pang, L., Lan, Y.Y., Guo, et al., J.F.: Text matching as image recognition[C], *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Arizona, USA, pp. 2793–2799 (2016)
- Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. [arXiv:1508.01991](https://arxiv.org/abs/1508.01991) (2015)
- Kim, Y.: Convolutional neural networks for sentence classification. [arXiv:1408.5882](https://arxiv.org/abs/1408.5882) (2014)
- Khodra, M.L., Wahyudi, A.S. Prihatmanto et al., Using graph-pattern association rules on yago knowledge base. In: *3rd International Conference on Science in Information Technology (ICS-ITech)*, IEEE, Bandung, Indonesia, pp. 136–141 (2017)
- Bromley, J., Guyon, I., LeCun, et al., Y.: Signature verification using a Siamese time delay neural network. *Advances in Neural Information Processing Systems (NIPS)*, pp. 669–688 (1993)

36. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. [arXiv:1409.2329](https://arxiv.org/abs/1409.2329) (2014)
37. Pennington, J., Socher, R., Manning, C.D: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532–1543 (2014)
38. Xu, B., Xu, Y., Liang, et al., J. Q.: CN-DBpedia: A never-ending chinese knowledge extraction system[C], In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Arras, France, pp. 428–438 (2017)
39. Li, P., Li, W., He, et al., Z.: Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. [arXiv:1607.06275](https://arxiv.org/abs/1607.06275), (2016)
40. nlp_chinese_corpus. Available online:https://github.com/FanniceCream/nlp_chinese_corpus
41. C. Zhang, L. Chang, W. Wang, H. Chen and C. Bin, Fine-grained question answering over knowledge graph based on BiLSTM-CRF[J], Computer Engineering, 46 (2019), 41–47
42. D. Luo, J. Su, P. Li, Multi-view attentional approach to single-fact knowledge-based question answering[J], Computer Engineering, 46 (2019), 215–221
43. S. Wang, J. Qiu, C. Hong, et al., Online commodity KBQA based on knowledge graph[J], Journal of Chinese Information Processing, 34 (2020), 104–112

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.