A Knowledge Graph-Based Question Answering System for Belt and Road Investment: Postprint

Authors: Chen Jinghao, Zeng Zhen, Li Gang

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] A knowledge graph-based Belt and Road investment question-answering system effectively integrates information resources from diverse sources, providing users with fast, accurate, and high-quality Belt and Road investment information, which holds significant research and application value.

[Method/Process] Information related to Belt and Road investment is collected, processed, and integrated to construct a Belt and Road investment knowledge graph under expert guidance. Based on this foundation, the functional components of the question-answering system are implemented, including: user question preprocessing, question classification, question template matching, and answer retrieval.

[Results/Conclusion] Experimental results demonstrate that the system can effectively answer Belt and Road investment-related questions.

Full Text

Preamble

A Question Answering System for "Belt and Road" Investment Based on Knowledge Graph

Chen Jinghao[1], Zeng Zhen[2], Li Gang[3]

[1]School of Public Policy and Management, Guangxi University, Nanning 530004
[2]School of Information, Guizhou University of Finance and Economics, Guiyang 550025
[3]Center for Information Resources Studies, Wuhan University, Wuhan 430072

Abstract: [Purpose/Significance] A question answering system for "Belt and Road" investment based on knowledge graph can effectively integrate informa-

tion resources from multiple sources, providing users with fast, accurate, and high-quality investment information about the "Belt and Road" initiative, which holds important research and application significance. [Method/Process] We collected, processed, and integrated investment-related information for the "Belt and Road" initiative, constructing a knowledge graph under expert guidance. On this basis, we implemented all functional components of the question answering system, including user question preprocessing, question classification, question template matching, and answer retrieval. [Result/Conclusion] Experimental results demonstrate that the system can effectively answer questions related to "Belt and Road" investment.

Keywords: question answering system; knowledge graph; Belt and Road; system construction
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.12.011

The "Belt and Road" initiative represents a crucial pathway for China to develop an all-around opening-up pattern in the new era [1]. Since General Secretary Xi Jinping proposed the initiative in 2013, Chinese enterprises have actively invested in countries along the route, with total direct investment exceeding $90 billion between 2013 and 2018, and newly signed overseas project contracts totaling over $600 billion [2]. As investment activities increase, so does the demand for information about host countries, investment environments, policies, and procedures. However, current approaches face two major challenges: first, relying solely on internet search engines to obtain "Belt and Road" investment information yields massive amounts of redundant data of varying quality, requiring substantial manual effort to extract useful knowledge; second, the multi-source, heterogeneous, and loosely structured nature of "Belt and Road" investment information resources results in poor integration and weak associations, making it difficult to provide standardized data and rich semantic expression.

Automatic question answering systems accept natural language questions from users, search for corresponding answers in knowledge bases, and return them to users [3]. Compared with traditional search engines, automatic question answering systems enhance the convenience of knowledge acquisition, save information screening time, and improve information quality. Traditional automatic question answering systems are mostly based on document retrieval, using keyword or template matching to query answers from unstructured text sources, which inherently suffers from insufficient query precision, reasoning capability, and semantic association. The emergence of knowledge graphs has changed this situation to some extent. Knowledge graphs represent knowledge bases that depict entities (concepts, people, things) and their relationships in the objective world in graph form [4], using triples as their representation format. Applying knowledge graph technology to automatic question answering systems helps extract structured knowledge from massive textual information, fuse data from different sources, and form a large-scale knowledge network rich in semantic relationships, thereby providing high-quality information for question answering

systems. By integrating knowledge graphs, the data precision, relevance, and structural level of question answering systems are significantly improved, enhancing the understanding and matching of question semantics and knowledge semantics. Based on this, constructing a knowledge graph-based question answering system for "Belt and Road" investment can address the aforementioned information acquisition problems.

In summary, we propose a design and implementation scheme for a knowledge graph-based question answering system for "Belt and Road" investment. This paper first reviews domestic and international research on question answering systems with brief commentary, then introduces the design concept and functional architecture of our system, followed by elaboration on the implementation process of key technologies, and finally presents experiments to demonstrate system usability and future work prospects.

2 Related Research Review

Automatic question answering systems can be categorized into open-domain and domain-specific systems based on their answer scope. Open-domain question answering systems are not restricted to specific domains and can answer questions across multiple fields, typically leveraging the redundancy of Web data resources and using statistical methods to find correct answers [5]. User questions in open-domain systems are relatively simple, using everyday language without scope limitations, and answers mainly come from Web resources [6]. These systems commonly use general semantic resources such as WordNet, HowNet, commonsense graph CYK, and linked data based on semantic web technologies like FreeBase and DBpedia [7]. Representative open-domain systems include the English question answering retrieval system AskJeeves, MIT's START [8], multilingual automatic question answering system AnswerBus [9], and IBM's Watson [10].

Domain-specific question answering systems generally only handle questions within a limited domain. Compared with open-domain systems, they process more professional and complex questions. Their target users are typically domain experts who use specialized terminology and demand high-quality answers. These systems are usually goal- and task-oriented, requiring domain knowledge bases and dictionaries that determine the scope of answerable questions. Due to their narrow domain and smaller user base, acquiring and building high-quality corpus resources is particularly valuable. Domain-specific systems have undergone long-term development, from structured data-based systems in the 1960s like Baseball [11] and Lunar [12], to computational linguistics-based systems in the 1970s-80s like Berkeley Unix Consultant [13], to free text-based systems in the 1990s, and FAQ-based systems at the beginning of this century, with continuous research achievements and technological progress.

Since Google launched its knowledge graph-based search product in 2012, the technology has been widely applied in artificial intelligence research, and knowledge graph-based domain-specific question answering systems have become

mainstream. Currently, many fields including life sciences, biomedicine, and library and information science have developed such systems. M. Vargas-Vera et al. developed AQUA, an academic domain question answering system where knowledge graph technology is used for query refinement, question reasoning, and similarity calculation [14]; A. Ben-Abacha et al. developed MEANS, a medical question answering system combining medical domain knowledge, natural language processing, and knowledge graph technology [15]; A. H. Asiaee et al. developed OntoLQA, a biomedical domain question answering system consisting of five main components: natural language processing, entity recognition, graph matching, semantic association, and answer retrieval [16]; X. Xie et al. built an automatic question answering system for a Natural Language Processing course, containing four processing modules: knowledge graph-based knowledge base, question analysis module, answer extraction module, and standard answer expansion module [17]; A. Abdi et al. established a physics domain question answering system using a semantic and syntactic information-based reasoning mapping method to transform user questions into knowledge base query languages [18]; A. Agarwal et al. constructed a dynamic conceptual network model integrating educational semantics, which improved the accuracy of the educational domain question answering system EDUQA [19]; Ma Chenhao created a thyroid knowledge graph and designed an automatic question answering system for thyroid diagnosis and treatment [20]; Cao Mingyu et al. constructed a primary liver cancer knowledge graph and implemented a pipeline question answering system [21]; Du Zeyu et al. proposed a streaming Chinese knowledge graph automatic question answering system CEQA capable of handling complex questions such as product consultation and statistical reasoning in e-commerce [22]; Lu Wei et al. built a library domain automatic question answering system based on Wuhan University Library's business requirements, introducing knowledge graph technology and establishing a multi-source data fusion knowledge base [23]. These studies elaborate on the construction process of knowledge graph-based automatic question answering systems from multiple perspectives and provide valuable references for this research.

Overall, as a support for the semantic web, knowledge graphs play a crucial role in automatic question answering and have become an effective way to organize, express, and manage massive, heterogeneous, and dynamic data. The "Belt and Road" initiative, as an important national development strategy, has received attention from government departments and research institutions, resulting in many websites, platforms, and databases such as the China "Belt and Road" website, "Belt and Road" channel, and "Belt and Road" database. These resources are valuable for guiding enterprises to invest in countries along the route. However, current utilization of these resources remains low, primarily at the text level without deep content exploitation, and information resource integration is insufficient. Therefore, by aggregating large amounts of "Belt and Road" investment-related information, creating a "Belt and Road" investment knowledge graph, and designing an automatic question answering system on this

basis, we can help users quickly, accurately, and comprehensively understand relevant knowledge and fill current research gaps.

3 System Framework

To implement a knowledge graph-based question answering system for "Belt and Road" investment, we must first address the acquisition and collection of data resources to support question answering, then process and organize these resources to form question answering corpus data. On this basis, we build the knowledge graph and construct the knowledge base. After the knowledge base is established, we further analyze and process user input questions, perform matching queries, and obtain final answers. Following this approach, our automatic question answering system can be divided into three major modules: data acquisition and processing module, knowledge graph construction module, and question analysis and answer retrieval module. The system framework is shown in Figure 1 [Figure 1: see original paper].

(1) **Data Acquisition and Processing Module**. This module includes information collection, format conversion, information filtering, information aggregation, and question clustering. Information collection gathers "Belt and Road" investment-related information from various data sources through web crawlers and manual methods, such as Q&A data from Baidu Zhidao and Zhihu platforms, the Ministry of Commerce's "Directory of Chinese Outward Investment Enterprises," "Directory of Foreign-invested Enterprises," and the "Investment Guide for Countries Along the Belt and Road" from the official China "Belt and Road" website. Format conversion transforms collected Excel spreadsheets and PDF documents into database tables. Information filtering removes redundant, noisy, and irrelevant information. Information aggregation integrates information from different sources—for example, the "Directory of Chinese Outward Investment Enterprises" only contains enterprise names without attributes like industry or region, which can be supplemented and integrated using enterprise information from Qichacha.com. Question clustering clusters questions from Baidu Zhidao and Zhihu to organize and summarize what netizens ask about "Belt and Road" investment on these platforms, providing references for subsequent question category division, knowledge graph construction, and question template creation.

(2) **Knowledge Graph Construction Module**. This module adopts a top-down approach for graph construction, including concept layer construction, instance layer construction, knowledge fusion, and knowledge base generation. Concept layer construction builds the "skeleton" of the knowledge graph by defining concepts, terminology, relationships, and attributes involved in the "Belt and Road" investment knowledge graph, clarifying its scope and standardizing its expression. The concept layer stores refined knowledge, typically managed using an ontology library that leverages axioms, rules, and constraints to regulate connections be-

tween entities, relationships, entity types, and attributes [24]. Instance layer construction builds upon the concept layer to extract entities, relationships, and attributes. Entity extraction, also known as named entity recognition, uses rule- and dictionary-based methods. Relationship extraction identifies associations between entities to form a knowledge network using dictionary-driven methods. Attribute extraction collects attribute information about entities from different data sources to achieve complete entity profiles, using the same approach as relationship extraction. After instance extraction, knowledge fusion methods organize the extraction results to eliminate contradictions and ambiguities through techniques including entity linking and knowledge merging. After knowledge fusion, facts are stored as triples in the form of "entity-relationship-entity" or "entity-attribute-attribute value," forming a graph-shaped knowledge network.

(3) **Question Analysis and Answer Retrieval Module**. This module includes question preprocessing, question classification, template matching, and answer query. Question preprocessing handles natural language questions input by users through the interface, including word segmentation, part-of-speech tagging, stop word removal, and entity recognition. Question classification uses automatic text classification technology to categorize processed user questions into predefined categories based on question types identified during the data acquisition and processing phase. This effectively reduces candidate answer space and improves the probability of returning correct answers. After classification, we use template-based matching [25] to understand questions. Question templates are designed based on common questions in each category and serve to map user questions to corresponding database query languages. The template matching process calculates similarity between user questions and prepared question templates using similarity algorithms. When the similarity value exceeds a certain threshold, matching is considered successful. When multiple templates exceed the threshold, the template with the highest similarity is selected. After template matching, the system understands question semantics based on identified entity names and relationship types, queries corresponding entities or attributes in the constructed "Belt and Road" investment knowledge graph, and returns query results to users as answers with coherent dialogue logic and grammatical fluency.

4 System Implementation Process

4.1 Data Acquisition and Processing

4.1.1 Q&A Data Acquisition Before constructing the "Belt and Road" investment question answering system, we first needed to collect domain knowledge, primarily through web crawlers and manual downloads. The web crawler module uses the HTML parser Jsoup, integrates the HTTPClient programming toolkit, and collects data from pages through

regular expressions. Collected content includes 25,034 entries from the Ministry of Commerce's "Directory of Chinese Outward Investment Enterprises" (http://femhzs.mofcom.gov.cn/fecpmvc/pages/fem/CorpJWList.html) and 3,180 entries from the "Directory of Foreign-invested Enterprises" (http://www.fdi.gov.cn/180000121_{10000207}8.html). Q&A data was crawled from Baidu Zhidao and Zhihu websites using keyword search patterns like "country name + investment" (e.g., "Singapore + investment"), yielding 31,555 question-answer pairs related to investment in countries along the route. To ensure knowledge authority, we also manually downloaded the "Investment Guide for Countries Along the Belt and Road" (https://www.yidaiyilu.gov.cn/info/iList.jsp?cat_{id}=10148) from the official China "Belt and Road" website. The system covers 64 countries along the "Belt and Road," categorized according to the classification scheme of Peking University's "Belt and Road" data analysis platform (see Table 1) [26]. The acquired data and downloaded investment guides cover these 64 countries. Ultimately, the system constructed 39,982 "Belt and Road" investment-related triples.

4.1.2 Q&A Data Processing After acquiring Q&A data, processing is required, including: converting country investment guide PDFs to text documents for storage (primarily using the PDFBox open-source software package); integrating data from the "Directory of Chinese Outward Investment Enterprises" and "Directory of Foreign-invested Enterprises" to form a "Belt and Road" enterprise investment database; using web crawlers to collect data from Qichacha.com to supplement enterprise attribute information such as region, industry, type, address, and business scope. Additionally, as references for subsequent knowledge graph and question template construction, collected Baidu Zhidao and Zhihu Q&A data must be filtered and clustered. Filtering mainly removes questions unrelated to "Belt and Road" investment, eliminates duplicates, and deletes null values. For clustering, we use the DBSCAN algorithm [27] for automatic cluster number determination, addressing the high-dimensional sparsity and semantic relationship deficiencies of traditional vector space models. Feature extraction uses Word2Vec combined with TF-IDF for text representation [28], with DBSCAN's eps parameter set to 0.5. The Word2Vec model was trained on 1.2 GB of Chinese Wikipedia corpus plus "Belt and Road" investment Q&A corpus from Baidu Zhidao and Zhihu, using the Skip-Gram model with 300-dimensional word vectors and a training window of 10. Automatic clustering yielded 2,240 question clusters, which after manual review, screening, and merging were reduced to 83 question categories containing 10,602 Q&A pairs. To ensure answer accuracy, we recruited five graduate students to review answers from Baidu Zhidao and Zhihu. Working in pairs, they selected accurate answers based on: (1) highest number of likes and answer time closest to system construction time, or (2) when like counts were low but timing was recent, students compared candidate answers (highest likes vs. most recent) and voted, with the answer receiving the most votes selected as the candidate answer.

4.2 Knowledge Graph Construction

Our knowledge graph uses a top-down construction approach, sequentially building the concept layer and instance layer. The concept layer primarily combines previously integrated data to extract and define terminology, concepts, and relationships, clarifying the overall scope. The instance layer populates data under concept layer constraints to ultimately form a structured knowledge graph. Below we describe the specific construction methods and storage approaches for the "Belt and Road" investment concept and instance layers.

4.2.1 Concept Layer Construction The concept layer design for the "Belt and Road" investment knowledge graph was built with domain expert guidance, combining knowledge from the "Investment Guide for Countries Along the Belt and Road," Baidu Zhidao, Zhihu Q&A, and the "Directory of Chinese Outward Investment Enterprises." To meet user questioning needs, we constructed both a "Belt and Road" country investment graph concept layer and an enterprise investment graph concept layer, including: domain concept induction and domain relationship and constraint definition.

(1) **Domain Concept Induction**. The country investment graph concept layer includes six core concepts: basic country information, basic investment information, investment laws and policies, investment procedures, investment precautions, and difficulty assistance. Basic country information refers to the profile of the host country, including four sub-concepts: national history, political environment, geographical environment, and social culture. Basic investment information reflects the investment potential of the host country, including seven sub-concepts: economic performance, domestic market, infrastructure, foreign trade and economy, financial environment, securities market, and business costs. Investment laws and policies refer to relevant investment regulations and policies of the host country, including 14 sub-concepts: foreign trade regulations, foreign investment market access, corporate taxation, foreign investment incentives, special economic zones, labor employment regulations, foreign enterprise land investment, foreign enterprise securities trading, environmental protection laws, anti-commercial bribery regulations, foreign enterprise project contracting, intellectual property regulations, investment cooperation laws, and business disputes. Investment procedures reflect how to handle investment formalities in the host country, including seven sub-concepts: investment registration, project contracting procedures, patent application, trademark registration, tax filing procedures, work permit processing, and investment consulting agencies. Investment precautions refer to situations requiring attention when investing in the host country, including five sub-concepts: trade precautions, project contracting precautions, labor cooperation precautions, risks to guard against, and other precautions. Difficulty assistance refers to ways to seek help when encountering difficulties in the host country, including four sub-concepts: seeking

legal protection, seeking government help, emergency plans, and protection from Chinese embassies. To clarify concepts and make them more specific, some third-level sub-concepts are further divided into fourth-level sub-concepts. For example, the third-level sub-concept "social culture" is further divided into ethnicity, language, religion, and customs. Ultimately, the country investment graph concept layer constructed six core concepts, 41 third-level sub-concepts, and 97 fourth-level sub-concepts, as shown in Figure 2 [Figure 2: see original paper].

Relative to the country investment graph, the enterprise investment graph is simpler, primarily serving Q&A about "Belt and Road" enterprise investment situations. It includes: investing country, home country, industry, type, address, registered capital, paid-in capital, and business scope as eight second-level sub-concepts, plus home region and investment region as two third-level sub-concepts, as shown in Figure 3 [Figure 3: see original paper].

(2) **Domain Relationship and Constraint Definition**. Relationships are core elements of the concept layer, describing interactions between concepts and instances in the domain and determining the richness of the knowledge graph. We primarily use two methods to define relationships between concepts: first, referencing high-quality data sources like the "Investment Guide for Countries Along the Belt and Road"; second, extracting existing relationship schemas from relational database tables like the "Directory of Chinese Outward Investment Enterprises." Ultimately, six major relationship categories were determined, as shown in Table 2 .

4.2.2 Instance Layer Construction After concept layer construction, the instance layer is built upon it. Instance layer construction primarily involves extracting "Belt and Road" investment knowledge matching the concept layer from previously acquired and processed records. This process handles both structured and semi-structured/unstructured data. The goal is to extract "Belt and Road" investment entities and relationships from different sources and represent them as triples. Specifically, instance layer construction includes entity extraction, relationship extraction, and attribute extraction.

(1) **Entity Extraction**. Based on concepts determined in the concept layer, we extract corresponding "Belt and Road" country names, enterprise names, etc., from data records to build entity nodes and form mappings from concepts to entities, such as country entities like Thailand, Vietnam, and Singapore, and enterprise entities like Meizhenxiang, Asia Pulp & Paper, and Zhongju Caifu.

(2) **Relationship Extraction**. Based on relationships determined in the concept layer, we build relationships between entities and determine relationship names according to inter-concept relationships. For example, if the concept layer defines an investment relationship between enterprises and countries, then based on the "Directory of Chinese Outward Invest-

ment Enterprises," we add investment relationships between enterprise instances (e.g., Wanda International Trade Group) and country instances (e.g., Singapore). Similarly, if the country basic information concept contains a social culture sub-concept, then at the instance level, Singapore's basic information and Singapore's social culture have a containment relationship. Additionally, if synonym relationships are defined in the concept layer, entities with synonym relationships are linked with their aliases.

(3) **Attribute Extraction**. Attribute extraction for the "Belt and Road" investment knowledge graph primarily relies on attributes contained in the concept layer corresponding to entities to extract attribute values. The "Investment Guide for Countries Along the Belt and Road" serves as a high-quality data source, allowing direct extraction of entity attributes and values. For example, Singapore's investment guide specifically introduces Singapore's customs, which can directly serve as attribute values for Singapore's customs attribute. In the "Belt and Road" enterprise investment database, Midea Group has the attribute "industry" with value "manufacturing," forming the triple <Midea Group, industry, manufacturing>. Table 3 shows partial mapping from concept layer to instance layer.

4.2.3 Knowledge Storage We use the popular open-source graph database Neo4j for knowledge graph storage. Implemented in Java, Neo4j stores structured data in network form. Compared with relational databases, Neo4j effectively addresses problems of low data value density and large data volume, providing a complete graph query language and supporting various graph mining algorithms. Neo4j provides Cypher statements for importing and querying data. For large-scale data, Neo4j also offers the neo4j-import tool for quickly importing large numbers of entities and relationships. We imported constructed "Belt and Road" investment-related triples into the Neo4j database using Cypher CREATE statements, Cypher LOAD CSV statements, and the neo4j-import tool. Figure 4 [Figure 4: see original paper] shows partial triple relationships from the "Belt and Road" investment knowledge graph in the Neo4j database.

4.3 Question Analysis and Answer Retrieval

After completing knowledge graph construction and storage, we can proceed with question analysis and answer retrieval. This involves preprocessing natural language questions input by users, classifying questions, obtaining computer query statements through template matching, and querying answers in the knowledge graph.

4.3.1 Question Preprocessing Question preprocessing primarily performs word segmentation, part-of-speech tagging, stop word removal, and entity recognition on natural language questions from users. We primarily use the HanLP open-source software package (https://github.com/hankcs/HanLP) for word

segmentation, which includes CRF-based word segmentation, part-of-speech tagging, and named entity recognition. We also built a custom dictionary consisting of country aliases and the "Belt and Road" enterprise directory to improve word segmentation and named entity recognition accuracy. Stop word removal mainly eliminates words with little practical meaning after segmentation, including modal particles, adverbs, prepositions, and conjunctions such as "的" (de), "在" (zai), and "啊" (a). After preprocessing, natural language questions input by users are abstracted in the system backend for subsequent classification and template matching. Table 4 shows examples of question preprocessing input and output.

4.3.2 Question Classification After question preprocessing, preprocessing results must be automatically classified into predefined question categories. In automatic question answering systems, question classification can effectively reduce candidate answer space and improve answer accuracy, while enabling different answer selection strategies for different question types. Combining processed Baidu and Zhihu Q&A data with the "Investment Guide for Countries Along the Belt and Road," we categorized common "Belt and Road" investment questions into six types: factoid questions, method questions, list questions, counting questions, judgment questions, and other questions. Factoid questions primarily answer "what is"; method questions answer "how to"; list questions query databases; counting questions perform statistical calculations on data meeting question conditions; judgment questions provide yes/no answers; other questions are those not belonging to the above five categories. Each category has typical characteristic words that guide classification. Table 5 shows the detailed classification.

After determining question categories, we converted questions in each category into abstracted examples and removed duplicates (see Table 5), resulting in 1,853 abstracted examples for automatic text classification. The classification algorithm used was SVM, with feature extraction using the same Word2Vec combined with TF-IDF approach as in question clustering [28]. We evaluated classification results using precision, recall, and F-measure, as shown in Table 6 . Classification results show that our text classification algorithm achieved a maximum F-measure of 96.53% and an average F-measure of 91.29%, demonstrating satisfactory performance for practical application.

4.3.3 Question Template Matching After question classification, user input questions are converted into corresponding templates for subsequent answer querying. Specifically, the question template matching process includes: (1) setting up corresponding template sets for common questions; (2) abstracting user input natural language questions and matching them with template sets to select the most similar template.

(1) **Question Template Design**. Based on the number and category of entities contained in questions, we designed a two-level, six-category question

template set with certain redundancy. The six categories correspond to question types from Section 4.3.2. The two levels are primary and auxiliary. Primary templates directly correspond to Neo4j graph database Cypher query statements and appear most frequently in user questions. Auxiliary templates share semantic meaning with primary templates to improve answer recall. Ultimately, we constructed 103 primary templates and 1,750 auxiliary templates. Table 8 shows template set examples.

(2) **Question Template Similarity Calculation**. After abstracting and automatically classifying user input natural language questions, we calculate similarity between processed user questions and templates in the template set. For similarity calculation, we first use Word2Vec and TF-IDF to convert questions into vectors [28], then use cosine similarity algorithm [29] to calculate similarity between user question vectors and template vectors. Through multiple experiments, we determined that when similarity between a question and template exceeds 0.75, the template is selected as the question template. When multiple templates exceed this threshold, the template with the highest similarity is chosen.

4.3.4 Answer Retrieval After obtaining the question template, the system uses the corresponding Cypher statement, combined with recognized entities and relationships, to query answers in the graph database and return them to users. The Cypher template for querying related entities with specific relationships is: `Match (a)-[:RelationName]-(b) where b.name='EntityName' return a.name`. Here, EntityName and RelationName are replaced with entity names and corresponding relationships recognized during question preprocessing in Section 4.3.1. For example, for the question "Which domestic enterprises have invested in Singapore?", after preprocessing, the system first recognizes the entity name "Singapore," then matches the template to obtain the corresponding relationship "investment," and finally embeds the entity name and relationship into the Cypher statement to query and obtain answers. Table 9 shows specific examples.

5 Experiments and Results Analysis

To test the accuracy of the "Belt and Road" investment question answering system, we designed 180 questions (30 per category) based on the classification in Section 4.3.2 to evaluate system performance. Answer accuracy is calculated as the ratio of correctly answered test questions to total test questions, as shown in Formula (1). The specific system operation process and experimental results are shown in Figure 5 [Figure 5: see original paper] and Table 10 .

The experimental results show that the system achieved an average answer accuracy of 81.1%, with most questions correctly understood and accurately answered. Although some questions used expressions inconsistent with system templates, the redundancy of our templates improved answer recall to some extent. Among question types, method questions achieved the highest accuracy,

while other questions had the lowest, with remaining types falling between these extremes. Analysis of incorrectly answered questions revealed that the system's semantic understanding capability needs improvement. For example, the system could not distinguish between "How many Chinese enterprises have invested in Singapore?" and "How many Chinese enterprises in Singapore have investment?" Although entities were correctly extracted, when entity order in the sentence was reversed, the system often returned incorrect answers. Additionally, as question templates increase, factoid questions and other questions are prone to confusion during text classification, causing answer matching errors. Future research will consider optimizing system modules to improve accuracy, including: (1) adding dependency parsing technology and deep learning to enhance question understanding; (2) expanding question coverage; (3) further improving and expanding the knowledge graph.

During testing, we also found that purely text-based answers were not ideal. Incorporating multimedia files such as images, videos, and audio might facilitate user understanding of answers. Additionally, domain-specific question answering systems are highly specialized and systematic, but users often don't know the system's domain boundaries and may ask irrelevant questions. Therefore, integrating the broad coverage and flexible answering methods of open-domain question answering systems into domain-specific systems represents a future breakthrough direction.

The maturation and deep application of big data, cloud computing, and artificial intelligence technologies have transformed traditional information services. New-generation information technology-supported services feature more flexible interaction, faster response, richer content, and mobile service delivery, bringing greater convenience while saving substantial labor costs. Therefore, this research represents preliminary exploration of intelligent information services and a beneficial attempt at smart information services, holding important practical significance and value. Based on knowledge graph technology, we constructed a "Belt and Road" investment question answering system. Under expert guidance and using publicly available data resources such as the "Investment Guide for Countries Along the Belt and Road," "Directory of Chinese Outward Investment Enterprises," Baidu Zhidao, and Zhihu Q&A, we established a "Belt and Road" investment knowledge graph and implemented system functions including question preprocessing, classification, template matching, and answer querying. Experiments demonstrate that the system can effectively answer "Belt and Road" investment-related questions. Future work includes further improving semantic understanding, expanding question coverage, and enhancing answer presentation capabilities.

References

[1] Xing Houyuan. Research on Investment Promotion under the "Belt and Road" Strategy [R]. Beijing: Investment Promotion Agency of Ministry of Commerce, 2017: 2.

[2] Wang Yuxiao, Yu Jiaxin. China's Direct Investment in Countries Along the "Belt and Road" Exceeds $90 Billion [EB/OL]. [2019-04-19]. https://www.yidaiyilu.gov.cn/xwzx/gnxw/86349.htm.

[3] Guo Tianyi, Peng Min, Yi Mulun, et al. Research Progress on Automatic Question Answering in Natural Language Processing [J]. Journal of Wuhan University (Natural Science Edition), 2019, 65(5): 417-426.

[4] Huang Hengqi, Yu Juan, Liao Xiao, et al. A Survey on Knowledge Graph Research [J]. Computer Systems & Applications, 2019, 28(6): 1-12.

[5] Brill E, Lin J, Banko M, et al. Data-intensive question answering [C]//TREC. Tenth Text Retrieval Conference. Gaithersburg: NIST, 2001, 56: 90.

[6] Wang Dongsheng, Wang Weimin, Wang Shi, et al. A Survey on Natural Language Understanding Methods for Domain-Specific Question Answering Systems [J]. Computer Science, 2017, 44(8): 1-8, 41.

[7] Lopez V, Miriam F, Motta E, et al. Poweraqua: supporting users in querying and exploring the semantic web [J]. Semantic web, 2011, 3(3): 249-265.

[8] Boris K, Gregory M, Gary B, et al. The start natural language question answering system [EB/OL]. [2019-09-10]. http://start.csail.mit.edu.

[9] Zheng Z. Answerbus question answering system [C]//Proceedings of the second international conference on human language technology research. San Francisco: Morgan Kaufmann Publishers Inc., 2002: 399-404.

[10] Ferrucci D, Levas A, Bagchi S, et al. Watson: beyond jeopardy! [J]. Artificial intelligence, 2013, 199(200): 93-105.

[11] Green Jr B, Bert F, Chomsky C, et al. Baseball: an automatic question answerer [C]//Proceedings of the western joint computer conference. New York: IRE-AIEE-ACM, 1961: 219-224.

[12] Woods W A, Kaplan R. Lunar rocks in natural English: explorations in natural language question answering [J]. Linguistic structures processing, 1977, 5(1): 521-569.

[13] Wilensky R, Chin D N, Luria M, et al. The berkeley UNIX consultant project [J]. Artificial intelligence review, 2000, 14(1/2): 43-88.

[14] Vargas-Vera M, Lytras M D. AQUA: A closed-domain question answering system [J]. Information systems management, 2010, 27(3): 217-225.

[15] Abacha A B, Zweigenbaum P. MEANS: a medical question answering system combining NLP techniques and semantic web technologies [J]. Information processing & management, 2015, 51(5): 570-594.

[16] Asiaee A H, Minning T, Doshi P, et al. A framework for ontology-based question answering with application to parasite immunology [J]. Journal of biomedical semantics, 2015, 6(1): 1-31.

[17] Xie X, Song W, Liu L, et al. Research and implementation of automatic question answering system based on ontology [C]//2015 27th chinese control and decision conference (CCDC). Piscataway: IEEE, 2015.

[18] Abdi A, Idris N, Ahmad Z. QAPD: an ontology-based question answering system in the physics domain [J]. Soft computing, 2018, 22(1): 213-230.

[19] Agarwal A, Sachdeva N, Yadav R K, et al. EDUQA: Educational domain question answering system using conceptual network mapping [C]//ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). Piscataway: IEEE, 2019: 8137-8141.

[20] Ma Chenhao. Design and Implementation of Automatic Question Answering System Based on Thyroid Knowledge Graph [J]. Intelligent Computer and Applications, 2018, 8(3): 102-107.

[21] Cao Mingyu, Li Qingqing, Yang Zhihao, et al. Knowledge Graph-Based Question Answering System for Primary Liver Cancer [J]. Journal of Chinese Information Processing, 2019, 33(6): 88-93.

[22] Du Zeyu, Yang Yan, He Hui. E-commerce Domain Question Answering System Based on Chinese Knowledge Graph [J]. Computer Applications and Software, 2017, 34(5): 153-159.

[23] Lu Wei, Qi Yue, Hu Xiaoge, et al. Design and Implementation of Library Automatic Question Answering System [J]. Information Engineering, 2019, 5(2): 5-16.

[24] Liu Qiao, Li Yang, Duan Hong, et al. Survey on Knowledge Graph Construction Techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.

[25] Unger C, Bühman L, Lehmann J, et al. Template-based question answering over RDF data [C]//Proceedings of the 21st international conference on world wide web. Lyon: ACM, 2012: 639-648.

[26] Wang Jimin, Wang Ruojia, Zeng Lanxin, et al. Evolution Analysis of Scientific Research Cooperation Networks in Countries Along the "Belt and Road" from 1996-2015 [J]. Library and Information Service, 2017, 61(16): 76-83.

[27] Zhang Xu, Sun Yuwei, Cheng Ying. Comparative Study on Effects of Different Features on Text Clustering: Taking News Text as an Example [J/OL]. Information Studies: Theory & Application: 1-13. [2019-10-19]. http://kns.cnki.net/kcms/detail/11.1762.G3.20190903.1330.006.html.

[28] Tang Ming, Zhu Lei, Zou Xianchun. A Document Vector Representation Based on Word2Vec [J]. Computer Science, 2016, 43(6): 214-217, 269.

[29] Gao Sen. Research on Question Classification and Similarity Calculation in Agricultural Question Answering System [D]. Hefei: University of Science and Technology of China, 2018: 32.

Author Contributions

Chen Jinghao: System design concept and paper writing
Zeng Zhen: System implementation and refinement
Li Gang: Paper concept determination and directional guidance

---

A Question Answering System for "the Belt and Road" Investment Based on Knowledge Graph

Chen Jinghao^{1}, Zeng Zhen^{2}, Li Gang^{3}

^{1}School of Public Policy and Management, Guangxi University, Nanning 530004
^{2}School of Information, Guizhou University of Finance and Economics, Guiyang 550025
^{3}Center for the Studies of Information Resources of Wuhan University, Wuhan 430072

Abstract: [Purpose/significance] The question answering system for "the Belt and Road" investment based on knowledge graph has important research and application significance as it can effectively integrate information from multiple sources and provide users with fast, accurate, and high-quality "the Belt and Road" investment information. [Method/process] We collected, processed, and integrated information related to "the Belt and Road" investment, and constructed the "the Belt and Road" investment knowledge graph under expert guidance. On this basis, the functions of each part of the question answering system were realized, including: question preprocessing, question classification, question template matching, and answer query. [Result/conclusion] The results show that the system can effectively answer questions about "the Belt and Road" investment.

Keywords: question answering system; knowledge graph; the Belt and Road; system construction

Note: Figure translations are in progress. See original paper for figures.