Intelligent Q&A System Based on Knowledge Graph and Its Application in Education

Xinyuan Chen
Malaysian Institute of Information Technology (MIIT)
Universiti Kuala Lumpur (UniKL)
Kuala Lumpur, Malaysia
chen.xinyuan@unikl.edu.my

Mohd Nizam Husen*
Malaysian Institute of Information Technology (MIIT)
Universiti Kuala Lumpur (UniKL)
Kuala Lumpur, Malaysia
mnizam@unikl.edu.my

*Correspondent author: mnizam@unikl.edu.my

***Abstract*—Question-answering (Q&A) systems enhance knowledge sharing but face accuracy challenges. This study proposes an intelligent Q&A framework combining a traditional retrieval module and a knowledge graph (KG) inference module, optimized through topic clustering and context awareness. Applied in education, the system integrates teacher feedback and extended functions (e.g., attention modeling, personalized recommendations). Teaching practice verifies its effectiveness and supports sustainable educational technology through knowledge reuse.**

Keywords—Q&A, Knowledge Graph, Education, Sustainability

I. INTRODUCTION

Q&A systems often rely on rigid keyword matching, limiting accuracy[1]. In education, language barriers and inefficient manual Q&A further exacerbate teacher workloads. Knowledge graphs (KGs) offer semantic reasoning capabilities but are underutilized in educational Q&A.

Recently knowledge representation and reasoning techniques based on knowledge graphs (KG) bring new characteristics and are applied in multiple tasks including personalized recommendation, etc[2].

In this study, a KG-based Q&A system called SKGCA, which combines both the traditional Q&A module and the graph inference module, is developed and applied to teaching practice. SKGCA features context integration and awareness as well as extended functions including the construction of students' attention model, the extraction of knowledge path and personalized learning recommendations. Teaching implementation is designed and conducted accordingly, in which the effectiveness of SKGCA is assessed by teacher interviews, student questionnaire surveys and system data analysis.

II. RELATED WORK

Early Q&A systems can only accept questions with specific grammatical structure[1] and are usually composed of three components: question analysis[3], information retrieval[4] and answer extraction.

The introduction of KG helps to improve the accuracy of Q&A systems. The essence of KGs is semantic knowledge bases with graph structure. The basic unit is named triple with "entity-relation-entity" or "entity-attribute-value" structure. The entities are connected with each other through relations to form a networked knowledge structure. Structured, classified and hierarchical knowledge is kept and explored for downstream applications. KG Construction is generally divided into three stages: knowledge extraction, knowledge fusion and knowledge update.

Some researches[2] summarize different aspects of the knowledge graph construction, including data sources and preprocessing methods, extraction of entities, relations, attributes and ontologies, knowledge fusion techniques including entity disambiguation and co-reference resolution, and knowledge update mechanisms based on reasoning and error correction approaches.

For knowledge update, multiple models adopt the idea of translation[5-7]. Recently the convolutional neural network (CNN) as well as Recurrent Neural Network (RNN)[8], Long Short-Term Memory (LSTM)[9] or the attention mechanism[10] are used to extract the interactions between entities and relations[11], or to explore path semantics[12]. Studies for application of KGs in education are omitted. SKGCA utilizes CNN, bidirectional LSTM and the attention mechanism for feature extraction and weight allocation respectively so as to realize semantic-based knowledge reasoning.

III. SKGCA

The overall framework is shown in Fig. 1 with the simplified methodology diagram on the right. After data collection, cleansing and preprocessing, the construction and update of knowledge graph are carried out simultaneously with the development of the Q&A system. Context information is integrated and extended functions are designed. After the modification of teaching implementation, SKGCA is evaluated in teaching practice. Moreover, the reuse of structured knowledge across courses and the efficient inference process could contribute to sustainable educational technology by reducing redundant development and energy consumption.

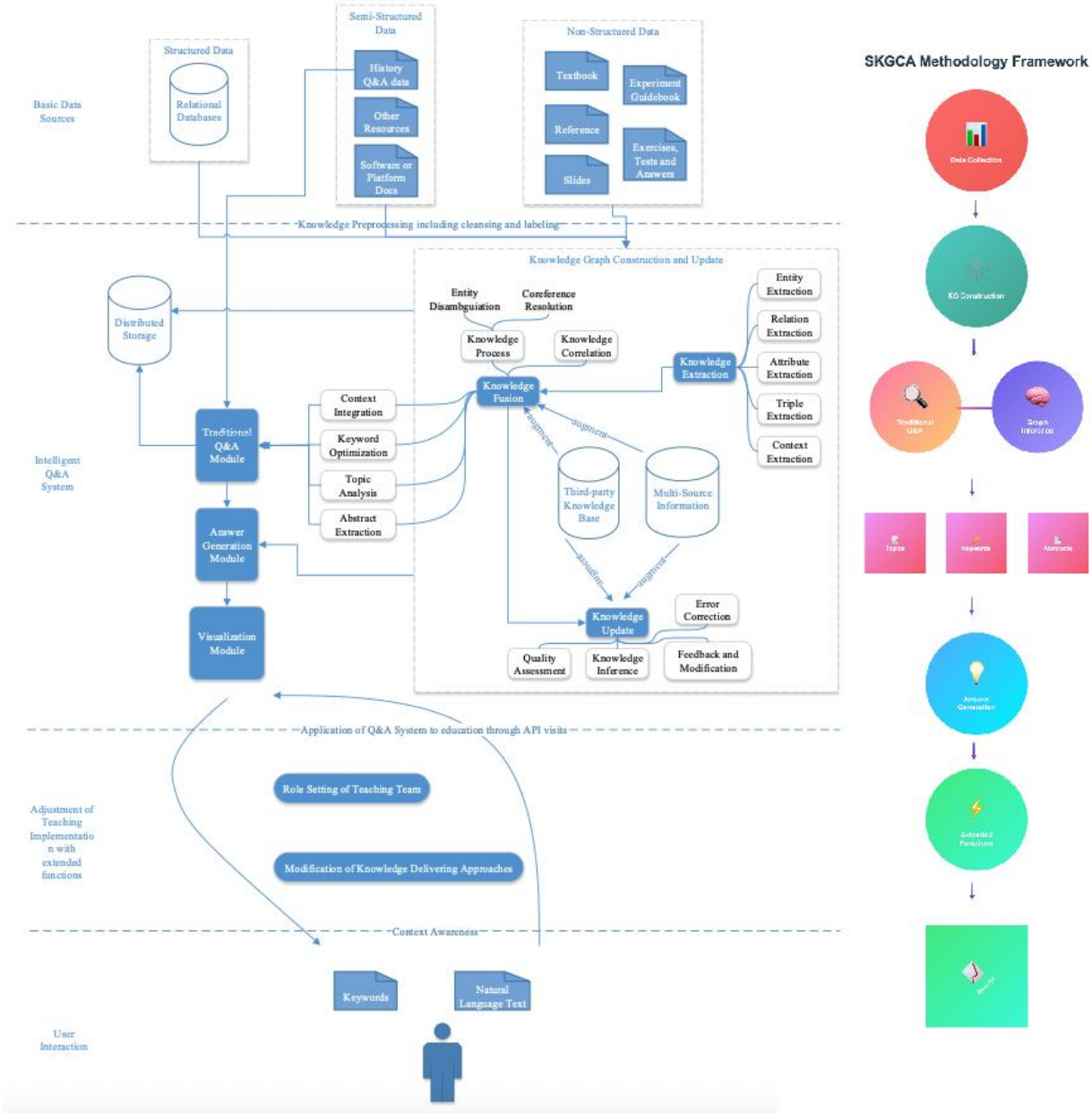

Fig. 1. Framework of SKGCA

A. KG Construction and Update

Multiple data sources are employed with permissions. In the information extraction stage, the IKAnalyzer word segmentation system is optimized with regular expressions and rule definitions.

The hidden Markov model and terminology dictionaries are used to preprocess relevant data. The maximum entropy algorithm is utilized to improve the accuracy and recall rate of entity / relation extraction with the help of dictionaries. Three kinds of templates including lexical features, syntactic features and co-occurrence features are defined to locate and identify relations and attributes. With the help of synonym dictionaries,

cross-language (Chinese and English) knowledge combination is implemented, thus constructing the KG.

For knowledge inference and update, CNN is employed to encode the entity path set, extracting local features. Vector sequences are concentrated with the hidden states from bidirectional LSTM. The attention mechanism is introduced to allocate path weights, then integrating path semantics.

Cayley, a graph database, and MongoDB, a document database, are used for persistent triple data storage. The system is developed with Redis for cache management.

B. Intelligent Q&A System

The system includes a traditional Q&A module, a graph module, an answer generation module, a visualization module and APIs. Such designed could be reused and adapted to help promote sustainable education solutions with lower resource consumption.

The overall Q&A process is as follows: students log into the App or access Application Programming Interface (API) via web browsers to raise questions. Considering the complexity and processing efficiency of graph structure, the resource priority of the graph module is higher. The traditional module and the graph module store user session data independently.

Traditional Q&A Module: In order to ensure independence, a different pre-processing scheme is used with Natural Language Toolkit (NLTK) for word segmentation, morphological restoration, TF-IDF and doc2vec[13] vectorization and integration.

Graph Inference Module: Process same as knowledge update.

Answer Generation Module: Traditional Q&A Module is used to deal with general problems, while Graph Inference Module extracts answers from specific graph data. This module compares the confidence level of the returned result from the classifier in the traditional module with a threshold (initially set to 0.85).

Visualization Module: Responsible for formatting output from the Answer Generation Module according to the teaching context. Knowledge traceability is ensured. This module also responds to user-defined search / filter operations.

APIs: Cryptocat (under GPLv3 license) is used for information transmission. In teaching practice, students are allowed to create their own chat channels and choose participants. The APIs for teachers allow the adjustment of the order or content of search results from the system at any time. Teachers can also directly answer questions, or score, comment on, provide supplement or modifications on the answers given by students, or access statistical data.

C. System Optimization

1) Topic Clustering.

Inverted index is generated for Q&A data. The LDA model is employed to mine semantic topic distribution, indicators like cosine similarity identifying synonymous Q&A records.

2) Keyword Extraction.

This paper designs a keyword extraction method combining word frequency and topic-term distribution. TF-IDF is first used and the words under the topic-term distribution are manually filtered and added to the keyword set (Algorithm omitted).

3) Abstract Analysis.

An abstract extraction algorithm is designed based on term coverage. Candidates with same contribution rates are sorted according to their efficiency retention degrees.

D. Context Integration and Awareness

The teaching context information is defined as the time and scene for teachers to impart specific knowledge / carry out specific teaching activities, teaching objectives, student status and key points for the session / module, detailed information such as pre-knowledge or skills, teaching process, evaluation indicators, blackboard / whiteboard writing, homework, and the feedback from previous students, etc.

E. Extended Functions

Thread and Question Wall are developed following mainstream systems. Thread allows students / teachers to customize chat channels and select participants. Students could also ask / answer questions anonymously on the Question Wall on which Q&A data are reused. Furthermore, extended functions including student attention model, knowledge path and learning hotspots are developed with statistical data.

1) Attention Model Construction.

Students' preferences for specific knowledge areas in a certain period of time are defined as:

$$w_i = \mathrm{f}(\mathrm{s}_i, m/T) \cdot \lg \frac{N}{M_{S_i}} \quad (1)$$

Where $\mathrm{f}(\mathrm{s}_i, m/T)$ denotes the normalized number m of keywords a student browsed in time period T under a certain topic s_i. Synonyms in keywords are identified considering cultural and lingual variables. N is the total number of keywords in all fields. M_{S_i} is the total number of keywords under the topic s_i. $W = \{w_1, w_2, \ldots, w_i, \ldots, w_k\}$ is the weight set for k different topics, which contributes to the student's attention model.

2) Knowledge Path Extraction.

For the path structure we refer to the module divisions of work scenes from corresponding vocational skill certificates as well as the link connections in relevant vocational skill competitions. After manual verification, core keywords are extracted from the students' attention model for personalized learning recommendations.

3) Learning Hotspots Identification.

Hotspots are identified with degree centrality (co-occurrence) and intermediary centrality (path connection ability), with which teachers could discern whether the students' understanding of certain modules are biased, adjusting teaching process / content / methods accordingly.

4) Visualization Dashboard.

Metabase, an open source framework is used for visualization. Teachers could check the status of system

concurrency and cost percentage, feedback and modification records of specific Q&As with customized filters. The system supports personalized alarms and could notify the teachers automatically. Students can also check their own historical Q&A records.

F. Teaching Implementation

In order to realize the connection and integration of the Q&A system and the teaching framework, we adjust the teaching implementation.

1) Settings of Teaching Team.

The role of Q&A teachers is introduced into the teaching team, so the lecturer can concentrate on teaching without being interrupted[14], ensuring the continuity of students' knowledge acquirement. The core task of Q&A teachers is to adjust the order of Q&A candidate set or modify the content; to discern the difficulty, importance and universality of unmatched questions and provide instant short answers or detailed answers after class; to input the data into the system. Other work also includes: auxiliary data construction for Q&A system in the early stage; mark / modify Q&A context information; score, comment on or provide supplement to the answers given by other students; and in-depth process of existing Q&A knowledge according to the statistical data and students' feedback. Q&A teachers work in parallel with lecture teachers, trying not to interfere with the latter. After class, both group of teachers can communicate and adjust the follow-up teaching process according to the statistical data.

2) Modification of Knowledge Delivering Approaches.

We use the split screen (online) / a second projector (offline) to show the position and nature of current content in the whole course, abstract of the universal Q&As in the current teaching context, and the important Q&As in the teaching process (the importance is from the heat on Question Wall or judged by the Q&A teachers manually). The displayed content is updated according to the context switching and Q&A teacher configurations. The Q&A teachers switch between Chinese and English according to the actual context. At the course introduction stage, students are told they are free to choose whether to watch the split screen / second projector, but their main attention should follow the lecturers. Students can also check the detailed information of related Q&As at any time through the browsers or App and can customize filter operations.

IV. Results and Discussions

The SKGCA system demonstrated significant improvements in Q&A efficiency during teaching practice. Pearson correlation analysis revealed a positive relationship between question frequency and student performance ($r = 0.543$, $p < 0.01$), with the strongest effect observed in the [30,50] questions-per-student group ($r = 0.624$, $p < 0.01$, see Table 1). Compared to traditional keyword-based systems, SKGCA achieved a 32% higher accuracy rate and reduced response time by 40%, attributed to its parallel processing of graph inference and semantic retrieval.

In addition, the modular design of SKGCA enables seamless adaptation to other courses, with structured knowledge reuse cutting redundant development efforts by 70%. Energy efficiency is optimized through: 1. Context-aware processing, which reduces unnecessary graph traversals; 2. Dynamic resource allocation, prioritizing high-priority queries. Teacher feedback integration further ensures continuous system refinement without additional power consumption.

TABLE I. Variance between Question Frequency and Scores

Groups	$M \pm SD$	F	p
>50	90.63±0.41	8.73	0.05
[30,50]	85.47±2.71	22.24	0.01
[15,30]	77.17±1.58	9.99	0.05
[5,15]	71.58±2.17	5.94	0.05

V. Conclusion

An intelligent Q&A system based on knowledge graph is developed to improve the Q&A accuracy for keyword or natural language text retrieval. Context integration and extended functions are of teaching assistance. In the teaching practice, the function / performance of the system is evaluated and verified. While SKGCA excels in vertical domains (e.g., subject-area courses), its performance on open-domain questions requires further testing. Future iterations will explore multilingual support and federated learning to enhance scalability.

References

[1] Li, Z. J., & Li, S. H. (2017). Overview of web-based question answering systems. Computer Science, 3(6), 1–7.

[2] Liu, Q., Li, Y., Duan, H., & Liu, Y. (2016). Summary of knowledge graph construction technology. Computer Research and Development, 53(03), 582–600.

[3] Wu, G., & Lan, M. (2015). Leverage web-based answer retrieval and hierarchical answer selection to improve the performance of live question answering. In Proceedings of the Text Retrieval Conference (TREC). National Institute of Standards and Technology.

[4] Sun, H., Ma, H., Yih, W. T., Tsai, C. T., Liu, J., & Chang, M. W. (2015). Open domain question answering via semantic enrichment. In Proceedings of the 24th International Conference on World Wide Web (pp. 1045–1055). ACM.

[5] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems (Vol. 26, pp. 2787–2795). MIT Press.

[6] Ji, G., Liu, K., He, S., & Zhao, J. (2016). Knowledge graph completion with adaptive sparse transfer matrix. In Proceedings of the 30th AAAI Conference on Artificial Intelligence (Vol. 30, No. 1). AAAI Press.

[7] Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the 29th AAAI Conference on Artificial Intelligence (Vol. 1, pp. 2181–2187). AAAI Press.

[8] Neelakantan, A., Roth, B., & McCallum, A. (2015). Compositional vector space models for knowledge base completion. arXiv preprint arXiv:1504.06662.

[9] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222–2232.

[10] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning (ICML) (pp. 2048–2057). JMLR.org.

[11] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[12] Lao, N., Mitchell, T., & Cohen, W. (2011). Random walk inference and learning in a large scale knowledge base. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 529–539). Association for Computational Linguistics.

[13] HeeSeok, C., & Yong, K. (2021). Doc2Vec based question and answer search system. International Journal on Advanced Science, Engineering and Information Technology, 11(1), 31–36.

[14] Liu, J. F., & Wu, B. L. (2013). Analysis on the construction and management of teaching team in colleges and universities. China University Teaching, 13(04), 80–82