An online question and answering system for support teacher-student interaction in the blended learning course

Viet Anh Nguyen, Manh Duy Nguyen
VNU – Universiy of Engineering and Technology, Hanoi, Vietnam
{vietanh, duynm_58}@vnu.edu.vn

***Abstract*— Student-teacher interaction plays a major role in teaching and learning in blended learning courses. During each lesson, students often do not have time to ask the instructor about the relevant issues of the subject that they want to answer. In addition, students often hesitate to talk directly with teachers about issues related to the content of the subject. The Question Answering (Q&A) system, therefore, is a useful tool to assist students and teachers to engage in interactive activities, especially in online courses. This research focuses on reviewing the state of the art of two respects as methods and techniques for developing question and answer systems, including issues of question classification, retrieval of information and extraction of answers as core components of the Q&A system. In addition, the paper proposes a model of an online questionnaire system for blended interactive learning activities which are being tested at a university training institution. Initially, the test results show that students are interested in this form of online question and answer.**

***Index Terms*— Q&A UET, question and answer system, blended learning, teacher-student interaction.**

I. INTRODUCTION

With the explosive development of information at present, the demand for searching and extracting information has become increasingly inevitable not only in social life but also in learning activities. Today, with the development of the Internet, people are able to communicate and work with extensive data and countless knowledge from many different areas of their lives. Accompanying with this development, there are higher requirements for the human capacity to manage, store and use knowledge resources from the collected data. Powerful searching systems such as Google, Bing, etc. allow people to search for data through various pieces of information. However, the data received from these searching engines is just the relevant material, and we still have to find the suitable answer to our needs ourselves. To find the most appropriate answer for the searching and finding information, the Q&A system is designed to give us the answer in a concise form instead of a package of related documents.

Based on the high applicability and the ability to solve the problems of data explosion, automated Q&A research has long been a concern of the scientific community all over the world. Specifically, from the 1960s, the first questionnaire systems using the database were born. Moreover, until the 1970s and 1980s, "textual understanding" was considered as a central issue to address in developing Q&A systems. It was also the first time that scientists have applied the natural language processing and analyzing model into the model of the Q& A system. At this stage, studies indicate that in order to have the right answers with the highest possible probability from the Q&A system, developers have to use a combination of methods relating to different areas. Among those, there are three most important areas: Natural Language Processing, Information Retrieval, and Information Extraction [1].

Also, along with the rapid development of the Internet, the demand for the online exchange of information in learning environments between learners and teachers is increasing, especially when the time spent in learning and teaching in classes is not enough for discussing and exchanging knowledge related to the subjects. Also, there is a lack of tools to assist the trainer to collect feedback or to directly make a quick assessment of the learner during the class. To meet this actual need, an interactive online Q&A system supporting the interaction in an integrated learning model is designed to facilitate the exchange and interaction of information between teachers and students.

II. LITERATURE REVIEW

In concept, Q&A is a system built with the task of finding the answer that fits the user-given question best. The question asked by the user and the answer of the system are both in the form of natural language containing information, which requires the system to be able to analyze the natural language [1].

In addition to the search engine, it is possible to view the automated inquiry system as a second option for users when they want to find the answer to a question in the form of a query. However, the main difference is the outcome of these two systems. While the search engine finds all documents containing information related to the keyword and the content of the query, the number of records returned may be huge when the question is unclear or inadequate. On the contrary, the Q &A system uses evaluation layers to return only one piece of text (or sentence) as the direct answer to the question.

In general, the automated question and answer system is divided into two main categories [2]:

• Closed-domain Question Answering System: This system is used to respond professional questions about a specific

Manuscript received August 7, 2017. (Write the date on which you submitted your paper for review.)

Viet Anh Nguyen is with the VNU – University of Engineering and Technology, E3, 144 XuanThuy, CauGiay, Hanoi, Vietnam (e-mail: vietanh@vnu.edu.vn).

Manh Duy Nguyen is with the VNU – University of Engineering and Technology, E3, 144 XuanThuy, CauGiay, Hanoi,Vietnam (e-mail: duynm_58@vnu.edu.vn).

area, such as education, health, transportation, sports, etc.

• Open-domain Question Answering: This is used to respond questions on all issues and areas without limits on question subjects and contents.

Because of its scientific and practical importance, the Q&A system has attracted the attention of many researchers from universities, research institutes and large companies in the information technology sector. Up to now, the design and construction of a Q&A system are no longer a new concept.

In the 1960s, the first automated questionnaire system was developed by some big names such as BASEBALL - the closed domain inquiry system developed by Green, Chomskey, and Laughery with the mission of answering questions related to the statistics of the American Baseball Federation [3]. In 1973, Woods developed the LUNAR system, a question and answer system similar to the BASEBALL system, to answer user's questions of stone samples returned from the moon exploration with the APOLLO probe. Systems developed during this period often seek answers by transforming the question from natural language into an available database query statement through analytical methods, mapping questions to data query language. After about ten years, Grosz developed the TEAM system[4] and made significant improvements over the above-mentioned two systems as he successfully applied the methods of analysis and semantic representation of the question by the first application of natural language processing (NLP) to a Q&A system. Regardless of the differences of used method and the working mechanism of these three systems, they have in common the use of knowledge databases for query and search purposes. These knowledge databases are widely hand-built by experts in their respective fields Under the achievements of the key natural language processing domain, open-domain question answering systems have been studied and developed over time with common names such as MARGIE - a system was studied by Schank, Goldman, Riesbeck, and Rieger in 1975 [5] and the START system (1997). These systems work by organizing documents containing information (another form of knowledge database) which are similar to the form of a human brain. They are also considered to be the earliest advanced questionnaire systems that operate as open domain Q&A systems do today.

The Q&A system is driven to develop not only by the growing demand for Q&A commercial systems but also by the emergence of large data sources such as WORDNET (Fellbaum, 1998), which provides information about semantic vocabulary and semantic links among groups of words in English . These free sources of data contribute to creating a more complex question and answer systems with more complex and accurate semantic processing.

In 2002, Li and Roth [6] introduced a method of classifying questions based on a machine learning approach to SnoW architecture first published by Khardon, 1999 with analytical features such as lexical words, part of speech tags, chunks, and named entities. They have used over 20,000 pre-labeled questions from various sources for training data and over 1000 unlabeled questions provided from the TREC conference as test data. The result of classifying questions of the system is not bad as it gets at almost 80% accuracy for over 50 different question classes.

Dell Zhang and Wee Sun Lee [7] introduced the classifying method using the SVM and compared performance with four other data exploring methods such as the K-neighbor nearest (KNN), Naïve Bayes (NB), Decision Tree (DT), Sparse Network of Winnows (SnoW). It uses 5500 test data and receives the following accuracy of algorithms: SVM reached 79.2%; DT reached 77%; SNOW reached 75.8%; NN reached 68.6%, and NB reached 67.8%, respectively. Thus, the approach to classifying questions by the vector-assisted algorithms achieves the highest efficiency compared to other methods.

In 2008, Zhiheng Huang et al. [8] introduced a method of classifying questions in the Q&A system with two approaches using a data mining model: Support Vector Machine SVM and Maximum Entropy Models (MEMs) on the basis of the vocabulary questions which are wh-question, head-word, semantic, N -grams and word blocks. Experimental results show that with the use of the word shape, the accuracy of the two approaches - The maximum support vector and Entropy model is 89.2% and 89% respectively over 50 different question layers; 93.4% and 93.6% respectively over 6 different question classes.

III. METHODS AND TECHNIQUES

This section summarizes some methods and techniques for developing a Q&A system focusing on natural language processing, searching, and extracting information.

In general, although these areas are not completely decisive to the architecture and quality of a Q&A system, it contributes to data processing and querying in the general architecture of the Q&A system. In particular, the Q&A system applies natural language processing methods to the analysis of questions and materials based on semantics and semantic relationships between vocabulary words. The search for information is a process in the Q&A system, with the task of finding all documents and information related to the content of the questions and ranking them. Extracting used information and analyzing the results of the information search help to find the most suitable answer for the user's question.

Based on specific applications from the three above-mentioned areas, the automated question and answer system is divided into the following three components [9]

A. Question Processing

This is a component that analyzes the question to find out the main focus of the question, the type of question and the appropriate type of answer. Question analysis plays an important role in any question and answer system. During this phase, questions are analyzed and processed to extract as much information as possible and can be used later in the data search phase. Specifically, the key issues that this component will implement are: (1) Analyzing the question to find out the essential content related to the question; (2) Classifying the question, then determining the appropriate type of answer; (3) Transforming questions from natural language into the query language.

In the processing of a Q&A system, the type of the key question is the main determinant of the searching method, evaluation model, and type of answer. Therefore, finding the right or wrong question will directly affect the outcome of searching the answer. Based on the type of question that the Q&A system serves users, questions are classified as follows

1) Entity, object question

This is the type of question about an object, entity, moment or manner and requires an answer in term of one sentence or a text. In English, questions of this kind usually start with words like what, where, when, how, or wh-question [2]. In Vietnamese, that sort of question is quite common. The entity needed for the answer can be analyzed, processed through the "entity labeling" system, and mapped directly into the source material. Also, this type of question does not require the system to analyze and manipulate natural language too complexly, without the need for logical reasoning. Disadvantages: The question posed by the user may be "Dim" due to general, unclear and insufficient content for finding the answer. And the processing of wrong semantic questions is a challenge to the question and answer systems in general.

2) List type questions

This type of question is used to ask for a list of entities and objects described in the question's text. Similar to the factoid type questions, the system easily detects the entity that needs to be queried through question analysis. Also, in order to answer questions of this type, the system does not require complex natural language processing. Determining and setting the threshold value (maximum number) for the query entity in the question is currently a processing issue of this type of question.

3) Hypothetical type questions

This type of question is used to ask for predictive information of a hypothesis, with the structure "What if ...?" Because it is hypothetical, this type of question is predictive and does not have the absolute correct answer [2]. This type of question-answering system can help users to analyze, judge the possibility of occurring or not for an event based on theory and data in practice. However, the accuracy and reliability of this kind of answering system are not high because sometimes the answer depends largely on the context of the question. It is not possible to apply common analytical methods, entity labeling for this type of question, which requires classification and clustering methods for predicting purposes.

4) Casual type questions

This type of question is used to ask about the reason for a fact or an event, usually, starts with the word "Why ..." or "Why not ...". Advantages: the answering system of this type of questions can help users to analyze, explain the phenomena, things that happen in reality. Disadvantage: However, to answer this question, the system uses a natural language processing and analysis system to find the main content of the question at the semantic and the contextual level.

5) Confirmation type questions

This type of question is used to require the system to confirm a given hypothesis and to require the system to respond "Yes" or "No" to the mentioned content or predictions. This type of answer has a simple answering syntax, without the need of constructing a separate answering system, only needs to apply the search results to figure out. However, to respond to this type of question, the system needs to have a mechanism of self-extraction and understanding of information from the data mining model.

B. Document Processing

Some information has been extracted during the question analysis phase such as the focus of the question, keyword, question type, answer type which will be used to search for information in the basic knowledge of the system. Many different methods can do this.

In particular, an open-domain Q&A system will use a search engine to search for documents distributed over the Internet. A closed-domain Q&A system that seeks information in data sources that is the sourced knowledge oriented to build by the expert in the area.

This main component task is to search information from various sources and then return the list of texts related to the content of the question; then the texts are filtered, selected and sorted by the level of relevance in order to choose the most appropriate text for the content of the answer. Specific functions include: (1) Searching for information: This is the first step of the component with the main function of finding information from various sources and finding information related to the questions through different evaluation criteria; (2) Filtering and selecting information: From the list of texts in the previous step, the system evaluates them, thereby refines, removes inappropriate texts or texts containing interfering content; (3) Sorting information: After finding a list of texts that match the question and possibly contains the answer, the system sorts them by logical levels so that they (can) become content that contains the correct answer.

1) Statistical-based approach

This method of data navigation is based on the notion of using quantitative properties to determine the probability of occurring the common vocabulary between the question and the paragraph, thus inferring similarities between them.

The necessary knowledge base for this approach is statistical probability, information theory, and linear algebra. This method does not require syntax or word processing and can be suitable for complex questions and heterogeneous data sources. However, this approach requires a sufficient amount of sample data to be used for computing, processing oriented data. It does not consider, handle the semantics and context of the sentence, or vocabulary.

2) Language-based approach

This method combines the use of language rules, knowledge and understanding to build models for question analysis, information retrieval, and document analysis. The system model is pre-built for both questions and answers and pre-tested for a sufficiently small scale and small error probability. The system does not need to be built on excessive sample data. It does not depend on the query language of the language-specific model and can be suitable with complex questions. However, the complexity, ambiguity in the syntactical structure and semantics of natural language of this method are some problems that need to be solved.

C. Answer Processing

Extracting the answer is a component that has an approach in the relevant field of information extraction. Inputs of

information in this phase may be documents or texts taken from the database query in the previous phase. The information will be used to extract the passage or semantically related sentence to the question.

In general, this component is responsible for identifying, constructing answers from lists of documents sorted and evaluated in the previous phase, and then checking for the found answers. Specifically, the key steps are: (1) Identifying "candidates" may be the answer to the question, based on the list of materials reviewed and sorted previously. (2) Extracting the answer by choosing words or phrases that can answer the question. The parser system allows finding the right content to answer the questions in paragraphs. (3) Evaluating the answer by calculating the reliability, the correctness, and the relevance. Some methods of evaluating the reliability of the answer are to use available vocabulary resources to check the correctness of the types of questions and answers.

Extracting the answer depends on the complexity of the question, the type of question found in the question analysis, the data that contains the content for the answer (searchable from the data source), the filtering method and contexts. So a question and answer system should be researched to extract the best answer from all of the above factors.

IV. A MODEL OF AN ONLINE QUESTION AND ANSWER SYSTEM

In this section, we introduce the QA-UET system - an open domain Q&A system in a learning model combined with questions posed by students. However, these questions will not be asked directly to the system but will be asked for the teacher who is teaching in the course. Thus, the online question and answer system in the learning model plays the role of an environment for the exchange of knowledge between the learner and the teacher. Also, the term "Q&A session" will be used by the system for teachers to handle student questions. It is considered as a small topic in which students' questions will be centered. Besides, organizing and managing students' questions in Q&A sessions help teachers to answer questions more easily and conveniently.

The questionnaire in the learning model has different characteristics from other questioning systems due to online requirements, which means that students' questions should be answered promptly by teachers during the course. Also, the questionnaire pattern of students has its own characteristics of diversity and does not follow a particular type of question on a given topic. In addition, the system allows students to search for similar questions that have been answered by reference. These are the main goals and directions for developing the Q&A system in the learning model.

For the Q&A system in the learning model, the peculiarity of the system is that there is no knowledge database including texts for finding answers; instead, the system considers questions answered by the teachers as the source of data used for finding questions similar to the student's questions. However, the answers to similar questions are not considered to be the final result for each question, but the teacher has to answer them. Therefore, the general architecture of the Q & A process is described as follows:

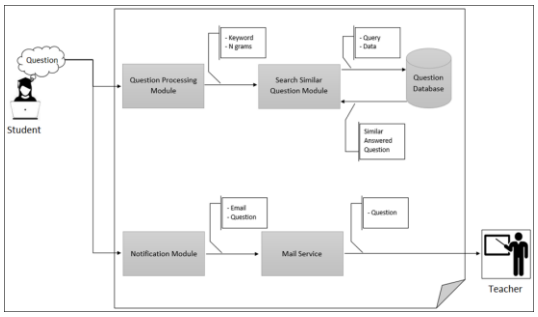

Fig 1. QA-UET system architecture

A. Processing operation of the system

Specifically, after the student asks a question with a particular subject of the teacher, the question will be recorded for the teacher to answer. Besides, for the convenience of student's reference, questions are analyzed to find similar questions that already have been answered in the system. Two steps to find the same question are "Question pre-processing" and "Question Matching." The question matching method is based on the N-grams features.

1) Preprocessing questions

A question posed to the system needs to be processed before comparing with available questions in the system in order to find out the similar question. With other question and answer systems, there are many steps in pre-processing and analyzing questions such as word separation, word exclusion, and keyword determination. However, within the framework of the paper, because the system can't use the word separation module and determine the word type, so in this step, the system only performs the removal of words that frequently appear but not carry essential meaning (known as stop word). After removing the stop words from the question, the remaining sentence is processed to separate the set of n-grams characters. A cluster of n-grams is a subsequence of n consecutive elements in a given sequence of elements. On the lingual context, the n-grams of a question is a set of n characters consecutively in that question. The n-grams feature is used to estimate the probability of a factor based on the factors surrounding it. In this study, we used the 2-grams (bigram) feature to analyze the question.

2) Matching the question

The main approach to finding the same question that was previously answered is to analyze the question in natural language to find a list of keywords or lists of vocabulary (N-grams). The system then uses the list of keywords or phrases found to generate a similar query search query in the database containing the previously answered question. The method of comparing two questions is based on the same (direct) number of keywords of those two questions. The ratio of the number of identical keywords to the total number of keywords for the two questions needs to reach a certain magnitude.

The inputs of the question matching step are two sets of 2-grams of the new question and one any question that is already answered in the system. The two questions are compared to the same degree by the Strike a match algorithm on the characteristic 2-grams syllable (this is a primary string-based method based on the number of occurrences of the features, so it does not consider factors regarding vocabulary or semantics). If consider the compared two sets of characteristic 2-grams syllables of the two questions as two

sets, the basic content of this method is to use the ratio of the number of identical elements of the two sets and the total element of two sets as evaluation parameters twice.

S = 2 * intersection / union * 100%

In it, the intersection is the number of the similar elements of the two sets and union is the total number of elements of the two sets. The two elements are the same when they contain two similar syllables and do not depend on the order.

B. Main functions of the system

The system includes a set of supporting functions that support the interaction between teacher - student, student - students. Teachers act as respondents in the Q&A system. The teacher manages students and student's questions through question-answering sessions, in which the Q&A session is a component of the QA-UET system created by the teacher to gather questions from participated students in the same course, subject area or field of study. Students act as "askers" in the QA-UET system. Students do not ask questions about the general subject. Each question is about the subject in a session class with course subject that the student is interested in. The main functions are illustrated in Figure 1.

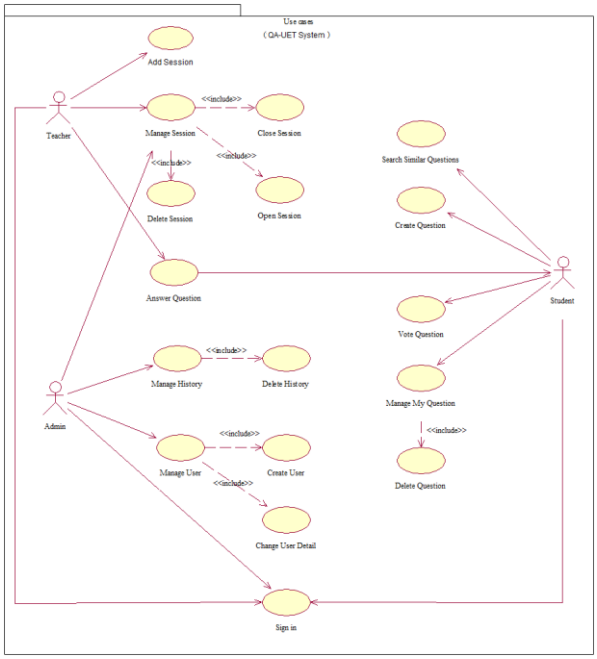

Fig.1. Main use case of Q&A System

1) Create a Q&A session

A session is a unit that stores questions about the same topic, the same field managed by the teacher. Each Q&A session will include the student's name, description, status, and a list of questions. The teacher acts as the respondent person answering questions in the inquiry system for this study. The teacher manages students and student's questions through Q&A sessions, where the Q&A session is a component of the QA-UET system created by the teacher to gather student's questions of the same subject, a topic or field of study. The teacher only has the authority to manage and respond to questions belong to his or her session. Students can ask questions at any question-and-answer session in the system without limits.

2) Create polls, surveys, assessments

This function allows teachers to create questionnaires with some common questions (in the current version, we support the following question types: yes/no, multiple choice questions with one answer; multiple choice questions with multiple answers) for online student surveys and assessing the level of understanding of knowledge during class.

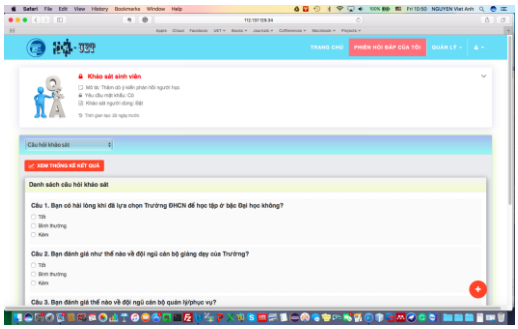

Fig.2. Student take survey in the class

3) Answer the question

The teacher selects "interesting" questions and answers; the teacher can choose how to respond directly to the class, or choose an answer from the answers of other students as a solution for the question.

4) Make a question

Students do not ask general questions to the system, but each question is placed in a question-and-answer session (of the teacher) in which the student is interested. To ensure the objectivity of each question, students can anonymously ask questions.

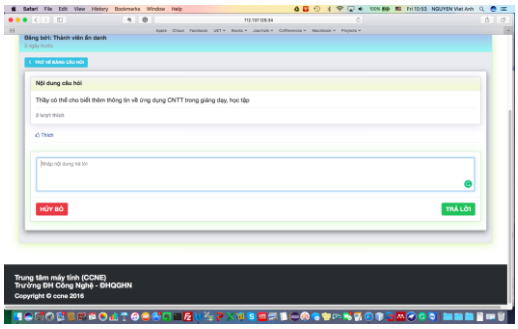

Fig.3. A interface of create question on the Q&A session

5) Comment, vote for the question, the survey

Students can answer questions from other students, vote for good questions and answers, answer surveys or student's lists of questions. Instructors can choose the best answer from the student's feedback as answers to that question.

V. DISCUSSION

Regarding the model, as our query language is Vietnamese, so there is some difficulties: (1) Vocabulary problem: Before processing, the question should be separated into words. However, unlike English, the vocabulary of Vietnamese is not simply separated by spaces because a word can contain many word components. Therefore, the separation of words needs to be done by a particular module. (2) The problem of word meanings: The richness of meaning in Vietnamese words makes a word to have different meanings or to be understood differently in each context. Therefore the question may be

difficult to understand if we do not know the context and the focus of the question. (3) Stop-word problem: The existence of words with high frequency but without significant meaning, similar to stop-words in English, such as connectors, pads, etc. may make the result diluted or redundant. Therefore, it is necessary to remove these words before performing a search.

The Q&A system UET also facilitates the creation of polls and surveys to collect quick feedback from learners during class activity. This helps the trainer to grasp the needs of the learner.

Regarding the system, nowadays there is a commercial product "Peagonhole" on the internet [10], which allows discussing question and answer sessions, to explore and to do online surveys. However, this system is not oriented for learning activities but focuses on seminars.

One of the limitations of our system is that there is currently no software version available for mobile devices. This inconvenience sometimes interferes learners in asking questions and give comments in class. This will be addressed by us in the next version.

VI. Conclusion

This study explored some methods and techniques for building an online questionnaire. Based on that, we have proposed a model and developed an interactive online Q&A system to support interaction in blended learning. The system makes it easy for students to have another way of effective interaction with the teacher while they participate in learning activities.

Although the model was recently tested, students welcomed it, and they were satisfied with the system .In the next stage, we will focus on issues related to the processing specific characteristics of the Vietnamese language in the query and answer selection. In addition, we will continue to develop a classify question module as well as answers recommended module so that the system can provide immediate answers to the students in case their question is a similar question database system.

VII. Acknowledgment

The authors wish to thank Dr. Nguyen Viet Ha, Msc. Do Hoang Kien have provided valuable comments to the research team for developing the Q&A UET system.

References

[1] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," Procedia Comput. Sci., vol. 73, pp. 366–375, 2015.

[2] [2] A. Mishra and S. K. Jain, "A survey on question answering systems with classification," Journal of King Saud University - Computer and Information Sciences, vol. 28, no. 3. pp. 345–361, 2016.

[3] [3] B. F. G. Jr., A. K. Wolf, C. Chomsky, and K. Laughery, "Baseball: an automatic question-answerer," in AFIPS Joint Computer Conferences, 1961, pp. 219–224.

[4] [4] B. J. Grosz, D. E. Appelt, P. A. Martin, and F. C. N. Pereira, "TEAM: An experiment in the design of transportable natural-language interfaces," Artif. Intell., vol. 32, no. 2, pp. 173–243, 1987.

[5] [5] R. Schank, G. Chris, R. I. Charles J, and R. Chris, "MARGIE MEMORY, ANALYSIS, RESPONSE GENERATION, and INFERENCE on ENGLISH," 1975.

[6] [6] X. Li and D. Roth, "Learning Question Classifiers: The Role of Semantic Information," Nat. Lang. Eng., vol. 1, no. 1, pp. 0–0, 1998.

[7] [7] D. Zhang and W. S. Lee, "Question classification using support vector machines," SIGIR '03 Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., pp. 26–32, 2003.

[8] [8] Z. Huang, M. Thint, and Z. Qin, "Question Classification using Head Words and their Hypernyms," Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP 2008), no. October, pp. 927–936, 2008.

[9] [9] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," Int. J. Res. Rev. Inf. Sci., vol. 2, no. 3, pp. 211–220, 2012.

[10] P. P. Ltd, "pigeonholelive." https://www.pigeonholelive.com/

Dr. Viet Anh Nguyen joined the VNU-University of Engineering and Technology in 2003. He defended his PhD Dissertation in 2010, after four years of study about adaptive hypermedia in e-learning. His research interests include e-learning, blended-learning, m-learning, user modeling, adaptive system, and recommender system. Viet Anh Nguyen can be contacted at: vietanh@vnu.edu.vn.

Manh Duy Nguyen is a student with VNU-University of Engineering and Technology. He is interested in computer technology enhanced learning. Manh Duy Nguyen can be contacted at: duynm_58@vnu.edu.vn