Using Syntax and Shallow Semantic Analysis for Vietnamese Question Generation

Phuoc Tran[1], Duy Khanh Nguyen[2], Tram Tran[3], and Bay Vo[4,*]
[1]Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology,
Ton Duc Thang University
Ho Chi Minh City, Vietnam
[e-mail: tranthanhphuoc@tdtu.edu.vn]
[2] Glarus, Axon Active Software Company
Ho Chi Minh City, Vietnam
[e-mail: khanh.nguyenduy@axonactive.com]
[3]Office of Science and Technology, Ho Chi Minh City University of Industry and Trade,
Ho Chi Minh City, Vietnam
[e-mail: tramtt@hufi.edu.vn]
[4]Faculty of Information Technology, HUTECH University,
Ho Chi Minh City, Vietnam
[e-mail: vd.bay@hutech.edu.vn]
[*]Corresponding author: Bay Vo

Received July 4, 2023; revised August 18, 2023; accepted September 12, 2023; published October 31, 2023

Abstract

This paper presents a method of using syntax and shallow semantic analysis for Vietnamese question generation (QG). Specifically, our proposed technique concentrates on investigating both the syntactic and shallow semantic structure of each sentence. The main goal of our method is to generate questions from a single sentence. These generated questions are known as factoid questions which require short, fact-based answers. In general, syntax-based analysis is one of the most popular approaches within the QG field, but it requires linguistic expert knowledge as well as a deep understanding of syntax rules in the Vietnamese language. It is thus considered a high-cost and inefficient solution due to the requirement of significant human effort to achieve qualified syntax rules. To deal with this problem, we collected the syntax rules in Vietnamese from a Vietnamese language textbook. Moreover, we also used different natural language processing (NLP) techniques to analyze Vietnamese shallow syntax and semantics for the QG task. These techniques include: sentence segmentation, word segmentation, part of speech, chunking, dependency parsing, and named entity recognition. We used human evaluation to assess the credibility of our model, which means we manually generated questions from the corpus, and then compared them with the generated questions. The empirical evidence demonstrates that our proposed technique has significant performance, in which the generated questions are very similar to those which are created by humans.

Keywords: Question generation; syntax-based; word segmentation; dependency parsing; named entity recognition.

http://doi.org/10.3837/tiis.2023.10.007 **ISSN : 1976-7277**

1. Introduction

Automated Questionnaire Generation (AQG) was defined by Rus et al. [1] as the task of automatically extracting questions from different representations including raw document, dataset, semantic representations, etc. AQG has many applications, such as: Intelligent Tutoring Systems (ITS), generating questions for the Stanford Question Answering Dataset (SQuAD) [2] and helping teachers prepare exam questions. In this research, we focus on generating questions from simple sentences, with the goal of preparing questions about travel and Nha Trang, which is useful for building the Vietnamese Question Answering (QA) dataset. For this task we primarily generated factoid questions, which means each question is generated from a single sentence.

In recent years, deep learning (DL) has been considered the most powerful tool within the artificial intelligence (AI) domain. It has numerous applications and is a promising direction for improving the performance of different tasks in NLP, including AQG [3]. However, in order to accomplish this task with acceptable performance, we need to have a massive volume of QA data, and for Vietnamese there is no large-scale corpus which can be used for the AQG task. Therefore, we cannot directly apply deep learning methods to perform question generation for this language.

Other approaches for automatically generating factoid questions from a sentence are subdivided into some strategies such as syntax, semantics, template, etc. The syntax strategy procedures follow a common approach: determining the syntactic form of the sentence, reducing the sentence's complexity (probably), recognizing main phrases, applying syntax conversion rules, and replacing investigation words. There are several syntax-based methods in the published research, with Kalady et al. [4], Varga et al. [5], Wolfe [6], and Ali et al. [7] providing examples thereof.

In the semantics-based approach, most proposed methods in this direction use transformations to convert sentences into questions. Recently, Mannem et al. [8] implemented a system that integrates Semantic Role Labeling (SRL) with syntactic transformations. With this a separate sentence is parsed with a semantic role identifier in the content selection phase to determine prospective targets. These targets are then chosen based on simple criteria. If present, all specific predicate semantic arguments (Argument 0 to Argument 5) are legitimate. It should be emphasized that a predicate with less than two of these arguments is deemed ineffective and ignored. In other words, it requires at least two arguments to sufficiently formulate the questions.

For the template-based approach, the techniques in this direction are mainly based on the concept that a question pattern could catch a group of specific context questions with a common form. Each of the many AQG strategies has advantages. While template-based approaches often create grammatically acceptable questions, syntax-based approaches cover the material more thoroughly. To provide more challenging human-like questions, the semantics-based strategies leverage some background resources such as named entity data, WordNet, Wikipedia, and so on. These types of questions have answers that are usually one or more words given. Some example questions for this category are: “Which continent does Vietnam belong to?” or “Who founded Microsoft company?”

In this paper, we mainly use a syntax-based approach to deal with the Vietnamese Question Generation (VQG) task, due to the limitations of existing Vietnamese NLP toolkits. Moreover, there are few useful studies that are related to this research area, such as SRL for Vietnamese. Specifically, to deal with the syntax-based approach, in the initial stage we need to have a thorough understanding of Vietnamese syntax. We then need to find as many structures of

sentences in Vietnamese as possible. Then, for each identified structure we find the target element and generate a general question format for each of these. To accomplish this, we need the support of some NLP toolkits that can carry out sentence segmentation, word segmentation, POS tagging, dependency parsing, and named entity recognition.

The structure of the rest of the paper is presented as follows: Part 2 gives main points of the works related to the problem of automatic question generation using syntax- and semantics-based approaches. Next, in Part 3, we formally present our proposed method for dealing with the VQG task. Then, the overall dataset description, experimental setups, experimental results and some discussions are presented in Part 4. Lastly, we end the paper by summarizing our attainment and future research in Part 5.

2. Background

2.1 Related work

DL is the main approach for problems using machine learning in recent times. DL-based methods have shown significant improvements in the performance of analytical and physical tasks without human intervention, including the AQG task.

In 2019, Kettip et al. [9] used Transformers trained with the SQuAD dataset [2] to deal with the AQG task in a machine reading comprehension (MRC) approach. To assess the efficiency of their model, the authors used the WER (Word Error Rate) to evaluate the similarity between the questions in SQuAD and the questions generated by their proposed model. The experimental results show that there was a significant difference between them. However, the questions generated using this technique are both correct in the grammar and semantic aspects. With the same approach to using deep learning, in 2021 Yuan et al. [10] proposed applying more pre-trained linguistic models to increase the quality of text features. These features were integrated into the sequence-to-sequence model [11] to guide the model to generate appropriate questions. This method was tested on two standard corpora and gave very promising results, with the BLEU score increasing by 17.2% and 6.2%, respectively, compared to previous baseline models. To gain the structural information in paragraphs, Xu et al. [12] proposed a sequence-to-sequence model with encoder applied to a Gated Recurrent Units network. The input to this network is a combination of components such as word embedding, answer tagging, and Graph Attention Networks embedding. Fei et al. [13] used the same approach, but they noticed that information is hidden in the previously generated text at each decoding step. To solve the problems, they designed the Iterative Graph Network-based Decoder to model the previous generation at each decoding step. Furthermore, their graph model catches dependency relations in the passage that raises the generation.

In the last few years, BERT [14] is a widely used deep learning architecture for many tasks in the field of NLP. In very recent times there have been two well-known studies within this direction, which are Chan et al. [15] and Alok Kumar et al. [16]. In the first [15], the authors used the Recurrent BERT model trained on the SQuAD corpus where the results illustrated that their suggested model substantially improved the performance of the AQG problem, with more than four scores in BLEU compared to previous models. In 2021, Kumar et al. [16] also used the BERT model to address the AQG problem, including two subproblems: AE (Answer Extraction) and QG. Specifically, the input of the model is a paragraph, and its outputs are some pairs of questions and answers. The generated questions are in the “Wh-” forms in English.

In 2021, Yu et al. [17] proposed a novel method that incorporates template-based QG model with a sequence-to-sequence model for diversity-aware QG. They did not apply stringent templates, instead they used adjustable patterns that can be collected effectively with less cost. The adjustable patterns cater high-level instruction of sentence structure while also enabling enough adjustability that the neuron-based model can fill chunks with content elements.

The literature shows that the deep learning approach is extremely effective for the question generation task when a massive amount of data is available. Currently, there are many question-answer corpora for rich resource languages including English, Chinese, French, etc. Nevertheless, there is currently no corpus that can be used for the AGQ problem for Vietnamese, and thus we cannot use a method based on deep learning to perform AGQ for this language.

Another approach to this task is a syntax-based method combined with a shallow semantic-based one. Keklik et al. [18] used semantic role labeling, dependency parsing, and named entity recognition to deal with generating questions. Unfortunately, such corpora are not available in the Vietnamese NLP community, and thus the use of PropBank and VerbNet as well as their related studies are beyond the scope of this paper.

In 2021, Bistak et al. [19] combined the traditional semantic and syntactic approaches with machine learning methods. First, the authors analyzed vocabulary, syntax, and semantic entities from the input, then proceeded to build a set of hierarchical rules for sentences. The set of features will be extracted from these samples, and then used for machine learning to generate transformation rules. In our method, we also use both syntax and semantics to create patterns. However, because of the lack of resources in Vietnamese, we have not combined the samples with machine learning algorithms, especially deep learning techniques.

We propose a novel AQG method based on syntax and shallow semantics, which are certainly not discrete. In addition, we mainly focus on generating factoid questions, a subject that has been studied by many researchers in the NLP area. There are some notable studies that apply the syntax-based approach, such as: Kalady et al. [4], Varga et al. [5], Wolfe [6] and Ali et al. [7]. These comply with a general approach: determining the syntax structure by parsing the sentence, reducing the sentence's complexity, recognizing basic phrases, applying syntax transformation constraints, and replacing question words. Using this strategy as the backbone, different authors have various approaches for each step.

- The first step is sentence simplification, which is considered a necessary pre-processing step. Each of the works mentioned above employ one or more simplification strategies. These include separating sentences with independent clauses, removing appositives, prepositional phrases, discourse markers, and relative clauses. One exception is Varga et al. [5], since the authors choose to apply a filtering process to remove complicated statements rather than simplify them.
- Key-phrase identification is the second step in the question generation syntax-based approach, and this means choosing phrases or single words which are considered as the best targets for question generation. Kalady et al. [4] used the methods of automatic summarization to recognize basic terms from a full text before jumping to sentence-level question generation. Another approach, presented in Ali et al. [7], is to identify a named entity in the subject, object, and prepositional phrases. Varga et al. [5] uses this approach and extends it to include other phrases, such as adverbials. To keep things relatively simple, the current study also applies key-phrase identification in order to choose answers to our questions.
- The final steps in syntax-based methods involve manipulating syntax trees and inserting interrogative words to convert declarative sentences into questions.

Heilman et al. [20] used established interaction rules to define in advance the varieties of questions created by SVO (Subject Verb Object) prepositional templates with specified named entity categories at each of these indexes. In [4], the authors used the previously mentioned method of identifying key terms at the document level to define separate heuristics. These heuristics are then utilized to produce questions based on whether the core phrases are present in object noun phrases, subject noun phrases, prepositional phrases, apostrophes, or adverbials.

2.2 Parsers

In this paper, we used Underthesea, a Vietnamese NLP toolkit [21]. The API library includes several powerful tools for handling Vietnamese text, including word segmentation, POS tagging, named entity recognition (NER), text classification, and dependency parsing. There are other tools available, some of which we will introduce in this section, for which the author provides a powerful API to utilize pre-trained NLP models to a Vietnamese text.

2.2.1 Vietnamese Syntax Rules

There are some main grammar rules in Vietnamese, as follows [22]:

- The overall sentence structure of Vietnamese is similar to that of English: Subject + Verb (+ Object) + Adverbs.
- There is no change in the form of the words, i.e., no inflection, and instead extra words are required to represent the increased meaning.
- When referring to particular things, nouns need to use a fit measure word.
- In Vietnamese, nouns come before adjectives.
- The present simple is similar to English.
- In the present continuous, Vietnamese uses the word “đang” (-ing in English) before verbs.
- The past and the future are expressed in the context or through time expressions.
- Adverbs are used in Vietnamese in a scenario similar to how they are used in English.
- In the negative, Vietnamese uses the word “không” (not) in places where English uses the words as don’t, doesn’t, isn’t, aren’t, etc.

2.2.2 Sentence Segmentation

Sentence segmentation is the process of identifying and separating sentences in a document by identifying where the punctuation is, and in Vietnamese this includes items such as: “.”; “!”; “?”. To do this we need to split the document based on the punctuation, but the rules must contain some exceptions, like email addresses, URLs, and proper nouns.

For example, we have a Vietnamese paragraph like “An sinh ra ở làng quê nghèo thuộc miền tây nam bộ Việt Nam. Năm 2015, anh ấy lên Sài Gòn học đại học và sau đó anh ta làm công việc Lập trình viên tại Sài Gòn sau khi tốt nghiệp. Hiện anh ấy là một trong những lập trình viên xuất sắc của công ty NUS” (An was born in a poor village in the southwestern part of Vietnam. In 2015, he went to Saigon to study at university and then he worked as a DEV in Saigon after graduating. Now he is one of the excellent DEV of NUS software company.). The paragraph us segmented into three separate sentences: (1) An sinh ra ở làng quê nghèo thuộc miền tây nam bộ Việt Nam. (2) Năm 2015, anh ấy lên Sài Gòn học đại học và sau đó anh ta làm công việc Lập trình viên tại Sài Gòn sau khi tốt nghiệp. (3) Hiện anh ấy là một trong những lập trình viên xuất sắc của công ty NUS.

2.2.3 Word Tokenization

Based on the previous findings on question generation in English, we considered that many points in the syntax-based approach can be applied to the Vietnamese question generation task. One of the main differences between English and Vietnamese is the word segmentation problem. That means we cannot just split text word by word based on the spaces, because in Vietnamese, for example, “tạm biệt” means “goodbye”, but if we split “tạm” and “biệt” then separately they will have different meanings, “tạm” that of “temporary” and “biệt” that of “separation”. We thus need some techniques for Vietnamese word segmentation that enable the mechanism to know which words can stand alone and which must be together, so that, for example, “tạm biệt” must be considered as one word, not two.

Dictionary-based approaches can also deal with this problem. These will scan the sentence left to right and recognize the words by looking them up in a dictionary. Additionally, the longest matching algorithm is needed to deal with the example mentioned above. where “Tạm” and “biệt” are already in the dictionary and have their meanings, but we need our model to know the longest word can be matched, which in this case is “tạm biệt”.

However, these traditional word segmentation methods do not give high performance due to the word ambiguity in Vietnamese. Recently, Wenkui Zheng et al. [23] released the most advanced deep neural network method that combined the Long-Short Term Memory (LSTM) and CNN models. The accuracy performance of this approach is 96.6%, the recall is 95.2%, and the F1 is 96.3%.

2.2.4 Part of Speech

Part-of-speech (POS) tagging is a common NLP task that involves recognizing words in a document that directly relate to a certain part of speech based on the meaning of the term and its context. In Vietnamese, one word may have multiple meanings, such as “đá” – in some cases its meaning is “stone” (noun) while in others it means “kick” (verb).

The model with the highest accuracy to deal with this problem is BiLSTM-CRFs, from Nguyen et al. [24], with an accuracy of 93.52%, based on a combination of bidirectional LSTM and Conditional Random Fields (CRF). This combination is mainly based on two levels of word information, such as (1) character-based word representations and (2) pre-trained word representations.

2.2.5 Chunking

Chunking (a.k.a. shallow parsing) is formally defined as a process related to NLP, and is normally used to identify POS and short phrases in a given sentence. There are several works which are based on POS tagging. However, instead of just supporting the identification and classification of a word’s POS (noun, verb, adjective, etc.), chunks provide clues about the sentence’s structure, the word’s structure or the phrase within a sentence (noun phrase, verb phrase, adjective phrase). For example, the Vietnamese sentence “bà ấy đang nấu ăn” (She is cooking) will be chunked into “NP[bà ấy] VP[đang nấu ăn] ”.

2.2.6 Named Entity Recognition

The process of finding and classifying entities named in the document is called Named Entity Recognition (NER). An entity can be a single/compound word that refers to the same object type (time, location, person, organization). NER plays an important role in the NLP area because it is a basic and indispensable task for handling multiple complex problems. For example, the Vietnamese sentence “Trần Tâm đang học tại Sài Gòn” (Tran Tam is studying in

Saigon) will be recognized as a named entity named "Trần_Tâm/PERSON đang học tại Sài_Gòn/LOCATION).

Many models have been proposed to deal with the Vietnamese NER task that have high F1 scores, but the highest score of 94.88% was obtained by the model in Nguyen et al. [24].

2.2.7 Dependency parsing

The technique of analyzing the grammatical structure of a phrase and detecting related terms as well as the kind of relationship between them is known as dependency parsing. The structure is defined by the relationship between a head word and its dependent words (children). A parse tree provides the most commonly used syntactic structure, and it allows navigating among the created parse trees that use head token and dependent ones. For example, the Vietnamese sentence "Tổng thống Obama đã đến thăm văn miếu tại Hà Nội." (President Obama visited the Temple of Literature in Hanoi) can be parsed as **Fig. 1.**

The VietTreebank of Nguyen et al. [25] s one of the useful resources used for Vietnamese dependency parsing.

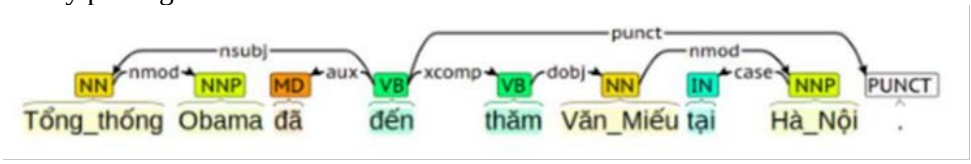

Fig. 1. An example of dependency parsing.

3. Proposed methods

3.1. The designed system

The system we designed for constructing the VQG can be divided into three stages: (1) the pre-processing stage, (2) the deconstruction stage, and (3) the construction stage.

- Pre-processing stage: We pre-process the sentence, which is the sentence simplification and word segmentation tasks.
- Deconstruction stage: We use a variety of semantic and syntactic parsers because these parsers describe their features of the sentence, and the more information we get, the better understanding we obtain. We will get the syntactic structure of a particular sentence by using dependency parsing. We then extract its POS information by using POS tagging, and by applying chunking to the pattern the system can then know the shallow syntactic structure, like a noun phrase, verb phrase, etc. With the shallow semantic analysis, we use the NER technique to identify which word refers to a person, location, time, or organization. The primary goal of the deconstruction phase is to obtain an intermediate representation in order to identify the sentence template, which is critical in classification of the generated question. The sentence pattern is the orderly arrangement of the roles of words in a sentence. "<Subject> + <Verb> + <Object> + <Complement>" is a good example of the sentence pattern.
- Construction stage: The designed system matches the sentence pattern with predefined rules. A question can be formed if a rule matches the sentence pattern. In order to ensure the final result is in the correct written form, it must be post-processed by adding, for example, a question mark and some interjection.

Fig. 2 illustrates the flowchart of the designed system.

3.2. Predefined Rules (Template)

Along with the Vietnamese Syntax Rules [22], we also obtained statistics on the paragraph field of two Vietnamese QA datasets [26][27] to find as many structures of sentences in Vietnamese as possible. These QA datasets only contain paragraphs for comprehension, and thus include a collection of facts or data that could be used to set some questions. Parsing these paragraphs gave us templates that identify which sentences can be used as the basis of questions.

This paper provides a rule-based strategy for generating questions from sentences. To be specific, both dependency parsing and named entity recognition techniques are used. Our contribution may be represented in terms of the following rules:

Dependency-based rules: We have some templates as follows:

- S-V-Obj: is a structure whose object predicate determines the subject.
- S-V-iobj: is an indirect object of a verb that receives the direct object.
- S-V-obl: is used for temporal and locational nominal modifiers.
- S-V-xcomp (open-clause complement): is a clause that does not have an internal subject.
- S-V-ccomp (clause complement): is a clause that has an internal subject.

NER-based rules include S-V-location, S-V-organization, and S-V-person.

There are currently 50 dependency-based and 10 NER-based rules in our approach.

3.3. An example

Fig. 3 shows the question generation process from the Vietnamese sentence “Tổng thống Obama đã đến thăm văn miếu tại Hà Nội.” (President Obama visited the Temple of Literature in Hanoi).

In this case, the input sentence does not change after going through the pre-processing step because our sentence does not include any abbreviations or special characters. In the next step, the sentence is put into the deconstruction stage which contains some parsers such as POS tagging, NER, chunking, and dependency parsing. With dependency parsing, “Tổng thống Obama” (President Obama) is the subject, “đã đến thăm” (visited) is the verb, and “Văn Miếu tại Hà Nội” (the Temple of Literature in Hanoi) is the object. Therefore, we can identify that it is in the S-V-obj form, and the system will generate questions about the object and subject. But which question word will be used – what, where, when, or who? The NER parser is used to deal with this problem, as it identifies which word is a person, location, organization, etc. Therefore, if the object is a person then the question word “who” will be chosen. In this example, our object (“Văn Miếu tại Hà Nội”) (the Temple of Literature in Hanoi) is tagged location. We then use the results of dependency parsing and compare with our defined rules. In this example, when the S-V-obj is matched, our questions will have two main formats: <Wh-V-obj>, and <Wh-aux-V-obj>.

4. Experimental results

4.1. Corpus

We used a corpus of more than one hundred simple sentences and a text about traveling in Nha Trang to experience our designed system. The document contains around 1,000 sentences and approximately 20,000 words. This text describes the nature, climate, culture, food, and places in Nha Trang, which is a coastal city located in the central part of Vietnam. According to the

corpus, there are 333 simple sentences and 79 complex sentences which show potential to generate questions.

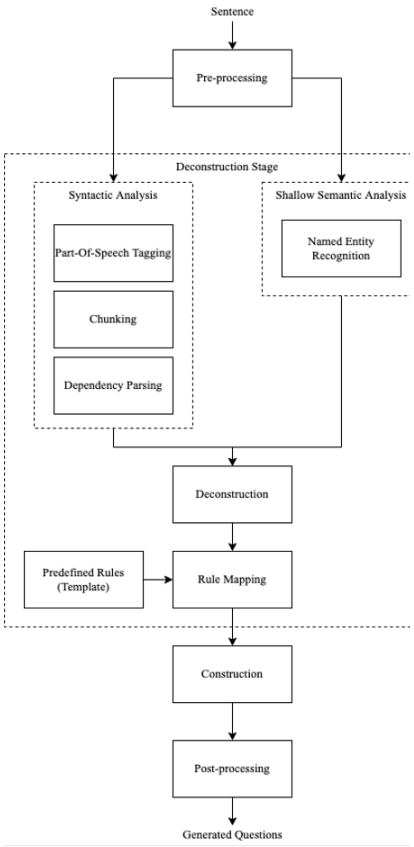

Fig. 2. Our designed system.

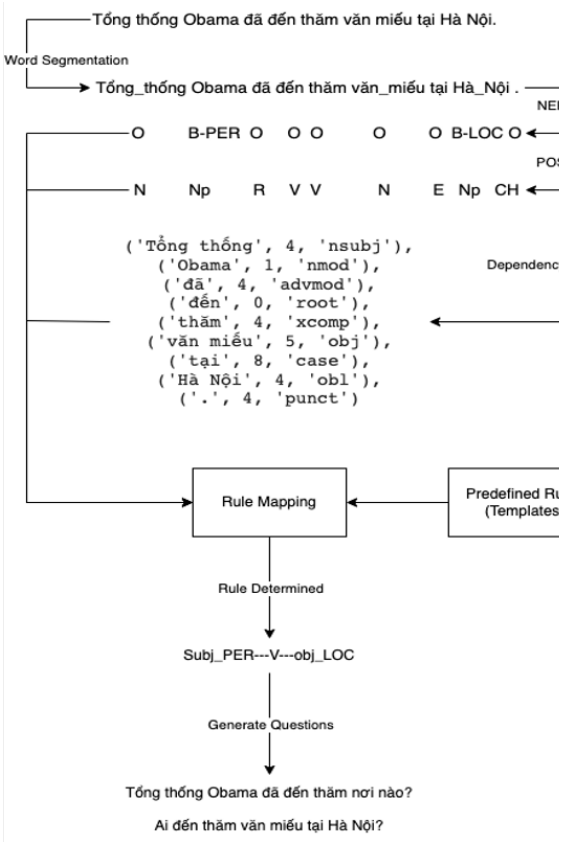

Fig. 3. An example of the question generation process.

4.2. Human evaluation

Human evaluation is required to evaluate the results of the QG model because questions can have several characteristics, such as being grammatical, making sense, or being vague. A question is considered grammatically correct if it conforms to the rules of grammar, while making sense and being vague are two disjoint categories – the former describes meaningful questions that are likely to be asked by humans, the latter represents invalid questions which would not be asked in the real world. **Table 1** shows the number of questions that were grammatically correct, made sense, or were vague for both simple and complex sentences.

Table 1. Evaluation of the results for 412 questions generated from the corpus.

Category	Simple (%)	Complex (%)
Grammatical	76.95	68.78
Make sense	52.31	42.31
Vague	47.69	57.69

There are three labels for each type of sentence: grammatically correct and make sense; grammatically correct but vague; grammatically incorrect and vague. **Table 2** gives an example for each label of question generated based on templates.

Table 2. Examples of labels from 412 generated questions.

Label	Sentence	Type of sentence	Question
Grammatical, make sense	Phần đất cũ của xứ Kauthara thuộc vương quốc Chăm Pa là Khánh Hoà ngày nay (Khanh Hoa today is the old land of Kauthara in the Champa kingdom.)	Simple	Đâu là phần đất cũ của xứ Kauthara thuộc vương quốc Chăm Pa? (Where is the old land of Kauthara in the kingdom of Champa?)
Grammatical, vague	Uỷ ban Quân quản Khánh Hoà chia Nha Trang thành 3 đơn vị hành chính: quận 1, quận 2 và quận Vĩnh Xương vào ngày 6 tháng 4 năm 1975. (On April 6, 1975, Khanh Hoa military administrative committee divided Nha Trang into 3 administrative divisions: District 1, District 2 and Vinh Xuong District.)	Simple	Uỷ ban Quân quản Khánh Hoà chia khi nào? (When did Khanh Hoa military administrative committee divide?)
Vague	Tỉnh Khánh Hoà được tái lập từ tỉnh Phú Khánh cũ vào ngày 1 tháng 7 năm 1989, Nha Trang là tỉnh lỵ tỉnh Khánh Hoà. (On July 1, 1989, Khanh Hoa province was re-established from the old Phu Khanh province, Nha Trang was the capital of Khanh Hoa province.)	Simple	Khi nào Nha Trang tái lập tỉnh Khánh Hoà? (When did Nha Trang re-establish Khanh Hoa province?)
Grammatical, make sense	Ngày 27 tháng 1 năm 1958, chính quyền Ngô Đình Diệm ban hành Nghị định 18 - BNV bãi bỏ quy chế thị xã, chia Nha Trang thành 2 xã là Nha Trang Đông và Nha Trang Tây thuộc quận Vĩnh Xương. (On January 27, 1958, Ngo Dinh Diem's government issued Decree 18 – BNV: abolished the town statute, divided Nha Trang into 2 communes including Eastern Nha Trang and western Nha Trang, belonging to Vinh Xuong district.)	Complex	Chính quyền Ngô Đình Diệm ban hành cái gì? (What did the Ngo Dinh Diem's government issue?)
grammatical, vague	Tiền thân của huyện Ninh Hoà ngày nay là Huyện Tân Định thuộc phủ Thái Khang, được hình thành có ranh giới từ đèo Rù Rì đến giữa sông Dinh. (Tan Dinh district belongs to Thai Khang metropolis, which is the precursor of Ninh Hoa district, formed with the boundary from Ru Ri pass to the middle of Dinh river.)	Complex	Huyện Tân Định thuộc phủ Thái Khang là gì? (What is Tan Dinh district of Thai Khang metropolis?)

Vague	Kể từ năm 1653, khi chúa Nguyễn Phúc Tần lập ra dinh Thái Khang với 2 phủ và 5 huyện, qua các đời vua Minh Mạng, Thành Thái, Duy Tân đến Bảo Đại sau này với nhiều cách phân định ranh giới, lập các đơn vị hành chính khác nhau. (Since 1653, when King Nguyen Phuc Tan founded Thai Khang Palace with 2 metropolises and 5 districts, through the reigns of Kings Minh Mang, Thanh Thai, and Duy Tan, Bao Dai with many ways of demarcating boundaries, setting up different administrative divisions.)	Complex	Khi chúa Nguyễn Phúc Tần lập ra dinh Thái Khang với 2 phủ và 5 huyện kể gì? (When King Nguyen Phuc Tan founded Thai Khang Palace with 2 metropolises and 5 districts, say what?)

In addition to the characteristics given above, we also had to compare the questions generated by the model with ones generated manually. We decided to use the F-score to measure this task with the following definitions:

$$\text{precision} = \frac{correct}{correct + spurious} \tag{1}$$

$$\text{recall} = \frac{correct}{correct + missed} \tag{2}$$

$$\text{F1} = 2 \times \frac{precistion \times recall}{precision + recall} \tag{3}$$

Here, correct refers to the number of questions generated by both the VQG model and manually. Spurious refers to the quantity of questions that are made by the VQG model but not manually. And missing mentions to the number of questions that are not created by the VQG model, but are generated manually.

Table 3 shows the number of questions which are correct, spurious, or missed and the F-scores for both two types of sentences.

Table 3. Evaluating the F-score for questions generated from simple and complex sentences.

	Correct	**Spurious**	**Missed**	**Precision**	**Recall**	**F-score**
Simple sentences	329	300	57	0.52	0.85	0.65
Complex sentences	77	105	39	0.42	0.66	0.51

4.3. Discussion

The first thing that struck us about the experimental results was that the generated questions were of quite good quality in terms of grammar and semantics. Selecting unknown components of a sentence and applying them to predefined patterns can be risky if their syntax and semantics are not compatible with the sentence content. As we mentioned, the designed system can generate many grammatically and semantically valid questions based on a few simple filters and modifications. However, the results would have been even more remarkable if the proportion for ambiguity had been smaller.

The F-score shows that the questions produced by our syntax-based system are quite different from those generated manually, especially the complex sentences. The reason is the vagueness of the Vietnamese templates, and thus our system only covered the two first levels of dependency parsing. In some cases, we must dissociate the third or fourth levels of the syntax tree to generate understandable questions. But this approach would be a challenge for us as computer scientists without collaboration with linguists. In the example "An chở mẹ đi chợ bằng xe máy" (An takes his mother to the market by motorbike.), our system could generate some questions thanks to the second level of the syntax tree, such as "An chở ai đi chợ bằng xe máy?" (Who is taken to the market by motorbike?), and "An chở mẹ đi đâu?" (Where does An take his mother?). This kind of question is better than the questions created by the first level of the syntax tree, like "An làm gì?" (What does An do?).

Moreover, the accuracy and reasonableness of the questions are proportional to the accuracy of dependency parsing and the NER technique. If a sentence has a wrong label from NER and dependency parsing, then the question which is generated will be wrong. For instance, if we take the sentence "New York là thành phố đông dân nhất của Hoa Kỳ." (New York is the most populous city in the United States.), and the NER used in this paper predicts New York is a person, then the question would be incorrect: "Ai là thành phố đông dân nhất của Hoa Kỳ?" (Who is the most populous city in the United States?).

5. Conclusions and Future Work

In this paper, we present a novel syntax-based technique for building a VQG system. Our proposed method is based on some parsers from the NLP field, such as sentence segmentation, word segmentation, POS tagging, chunking, dependency parsing, and NER. In our approach, both sentence and word segmentation are used for pre-processing. Then in the deconstruction stage we use syntactic analysis combined with shallow semantic analysis to understand the structure as well as the sense of the sentence. After that we apply some predefined templates to match the extracted structure. A single sentence may produce one or more questions, as we can ask about the subject, predicate, time, location, etc.

One limitation of our model is that the questions generated are factoid questions, which means that students only need to remember the information in a text and may not necessarily understand it. Moreover, factoid questions are usually not very valuable in examinations or for students to evaluate their own knowledge. We thus plan to extend our work by researching other techniques for generating questions from multiple sentences at the paragraph level, and to Transformers, which is the state-of-the-art model in NLP.

References

[1] Vasile Rus, Zhiqiang Cai, and Art Graesser, "Question generation: Example of a multi-year evaluation campaign," in *Proc. of QGSTEC*, 2008. Article (CrossRef Link)

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing, ACL*, pp. 2383-2392, 2016. Article (CrossRef Link).

[3] Ying-Hong Chan, Yao-Chung Fan, "BERT for Question Generation," in *Proc. of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics*, pp. 173-177, 2019. Article (CrossRef Link).

[4] Saidalavi Kalady, Ajeesh Elikkottil, and Rajarshi Das, "Natural language question generation using syntax and keywords," in *Proc. of QG2010: The Third Workshop on Question Generation*, pp. 1-10, 2010. Article (CrossRef Link).

[5] Andrea Varga and Le An Ha, "Wlv: A question generation system for the QGSTEC 2010 task B," in *Proc. of QG2010: The Third Workshop on Question Generation*, pp. 80-83, 2010. Article (CrossRef Link)
[6] John H Wolfe, "Automatic question generation from text-an aid to independent study," *ACM SIGCUE Outlook*, vol. 10, pp. 104-112, 1976. Article (CrossRef Link).
[7] Husam Ali, Yllias Chali, and Sadid A Hasan, "Automation of question generation from sentences," in *Proc. of QG2010: The Third Workshop on Question Generation*, pp. 58-67, 2010. Article (CrossRef Link).
[8] Prashanth Mannem, Rashmi Prasad, and Aravind Joshi, "Question generation from paragraphs at UPenn: Qgstec system description," in *Proc. of QG2010: The Third Workshop on Question Generation*, pp. 84-91, 2010. Article (CrossRef Link).
[9] Kettip Kriangchaivech and Artit Wangperawong, "Question Generation by Transformers", *arXiv:1909.05017*, 2019. Article (CrossRef Link).
[10] Wei Yuan, Tieke He, and Xinyu Dai, "Improving Neural Question Generation using Deep Linguistic Representation," in *Proc. of the WWW Conference 2021*, pp. 3489-3500, 2021. Article (CrossRef Link).
[11] Ilya Sutskever, Oriol Vinyals, and Quoc Le Viet, "Sequence to Sequence Learning with Neural Network," *arXiv:1409.3215*, 2014. Article (CrossRef Link).
[12] J. Xu, Y. Sun, J. Gan, M. Zhou and D. Wu, "Leveraging Structured Information from a Passage to Generate Questions," *Tsinghua Science and Technology*, vol. 28, no. 3, pp. 464-474, 2023. Article (CrossRef Link).
[13] Zichu Fei, Qi Zhang, and Yaqian Zhou, "Iterative GNN-based Decoder for Question Generation," in *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2573-2582, 2021. Article (CrossRef Link).
[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171-4186, 2019. Article (CrossRef Link).
[15] Ying-Hong Chan and Yao-Chung Fan, "A Recurrent BERT-based Model for Question Generation," in *Proc. of the 2nd Workshop on Machine Reading for Question Answering*, pp. 154-162, 2019. Article (CrossRef Link).
[16] Alok Kumar, Aditi Kharadi, Deepika Singh, and Mala Kumari, "Automatic question-answer pair generation using Deep Learning," in *Proc. of 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 794-799, 2021. Article (CrossRef Link).
[17] Xiaojing Yu and Anxiao Jiang, "Expanding, Retrieving and Infilling: Diversifying Cross-Domain Question Generation with Flexible Templates," in *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3202-3212, 2021. Article (CrossRef Link).
[18] Onur Keklik, Tugkan Tuglular, and Selma Tekir, "Rule-Based Automatic Question Generation Using Semantic Role Labeling," *IEICE Transactions on Information and Systems*, VOL. E102–D, NO. 7, pp. 1362-1373, 2019. Article (CrossRef Link).
[19] Miroslav Bistak and Viera Rozinajova, "Automatic question generation based on sentence structure analysis using machine learning approach," *Natural Language Engineering*, Vol. 28, no. 4, pp. 487 – 517, July 2022. Article (CrossRef Link).
[20] Michael Heilman and Noah A Smith, "Extracting simplified statements for factual question generation," in *Proc. of QG2010: The Third Workshop on Question Generation*, pp. 11-20, 2010. Article (CrossRef Link).
[21] Anh Vu, "Underthesea – Vietnamese NLP toolkit," *https://pypi.org/project/underthesea/*, 2018. Article (CrossRef Link).
[22] Thu Dao Minh, Ngoc Dao Thi Minh, Van Nguyen Mai, Ngan Le Kim, Huong Le Thanh, Thai Nguyen Phuong, Lam Do Ba, "Vietnamese Syntax Rules," *Association for Vietnamese Language and Speech Processing*, 2009. in Vietnamese, [Online]. Available: http://www.jaist.ac.jp/~bao/VLSP-text/Mar2009/SP85_baocaokythuat2009thang3.pdf

(accessed in February 28, 2023).
[23] Kexiao Zheng and Wenkui Zheng, "Deep Neural Networks Algorithm for Vietnamese Word Segmentation," *Scientific Programming*, 2022. Article (CrossRef Link).
[24] Duong Nguyen Anh, Hieu Nguyen Kiem, and Vi Ngo Van, "Neural sequence labeling for Vietnamese POS tagging and NER," in *Proc. of 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2019. Article (CrossRef Link).
[25] Thai Nguyen Phuong, Luong Vu Xuan, Huyen Nguyen Thi Minh, Hiep Nguyen Van, and Phuong Le Hong, "Building a large syntactically-annotated corpus of Vietnamese," in *Proc. of the Third Linguistic Annotation Workshop on ACL-IJCNLP*, pp. 182-185, 2009. Article (CrossRef Link).
[26] Kiet Nguyen, Vu Nguyen, Anh Nguyen, Ngan Nguyen, "A Vietnamese Dataset for Evaluating Machine Reading Comprehension," in *Proc. of the 28th International Conference on Computational Linguistics*, pp. 2595-2605, 2020. Article (CrossRef Link).
[27] Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen, "New vietnamese corpus for machine reading comprehension of health news articles," *ACM Transactions on Asian and Low-Resource Language Information Processing*, VOL. 21, NO. 5, 1-28, 2022. Article (CrossRef Link).

Phuoc Tran received the B.S. degree in information technology from the University of Pedagogy, Ho Chi Minh City, Vietnam, in 2006, and the M.Sc. and Ph.D. degrees in computer science from the VNU Ho Chi Minh City University of Science, in 2011 and 2018, respectively. He is currently a Researcher with Natural Language Processing and Knowledge Discovery Laboratory, Ton Duc Thang University, Vietnam. His research interests include natural language processing, machine translation, question answering, text mining, and database systems.

Duy Khanh Nguyen is a research student at Ton Duc Thang University since 2020, as well as a programmer at Axon Active Vietnam software company. He graduated with a Bachelor's degree with Honors in May 2023 and achieved many achievements in school-level scientific research competitions on Natural Language Processing. His research topic is related to language processing, question answering systems, question generation systems, and database systems.

Tram Tran received the B.S. degree in Information Technology and M.Sc in Computer Sci ence from the University of Information Technology, Vietnam National University, in 2009 and 2017, respectively. She is currently working and researching at Office of Science and Technology, Ho Chi Minh City University of Industry and Trade, Vietnam. Her research interests include association rules, Text mining, Question Answering, and database systems.

Bay Vo received the Ph.D. degrees in computer science from the University of Science, Vietnam National University, Ho Chi Minh City, in 2011. He is currently an Associate Professor and Vice President of HUTECH University, Ho Chi Minh City. His research interests include data mining, text mining, and social network analysis.