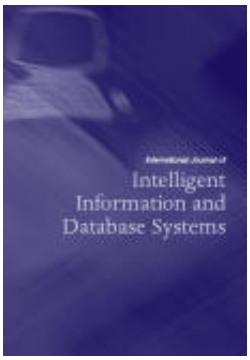


International Journal of Intelligent Information and Database Systems

ISSN online: 1751-5866 - ISSN print: 1751-5858
https://www.inderscience.com/ijiids

A collaboration of an ontology and an autoregressive model to build an efficient chatbot model

Thi Thanh Sang Nguyen, Dang Huu Trong Ho, Ngoc Tram Anh Nguyen, Pham Minh Thu Do

DOI: 10.1504/IJIIDS.2023.10059348

Article History:

Received:	21 August 2022
Last revised:	12 March 2023
Accepted:	07 August 2023
Published online:	02 April 2024

A collaboration of an ontology and an autoregressive model to build an efficient chatbot model

Thi Thanh Sang Nguyen*,
Dang Huu Trong Ho,
Ngoc Tram Anh Nguyen and
Pham Minh Thu Do

School of Computer Science and Engineering,
International University,
VNU-HCMC; Vietnam National University,
Ho Chi Minh City, Vietnam
Email: nttsang@hcmiu.edu.vn
Email: hdhtrong@hcmiu.edu.vn
Email: nntanh@hcmiu.edu.vn
Email: dpmthu@hcmiu.edu.vn
*Corresponding author

Abstract: Question-answer systems are now very popular and crucial to support human in automatically responding frequent questions in many fields. However, these systems depend on learning methods and training data. Therefore, it is necessary to prepare such a good dataset, but it is not an easy job. An ontology-based domain knowledge base is able to help to make nice question-answer pairs and reason semantic information effectively. This study proposes a novel chatbot model involving ontology to generate efficient responses automatically. Besides, an autoregressive model is also employed to complement responses. A case study of admissions advising at the International University – VNU HCMC is taken into account in the proposed chatbot. Experimental results have shown that the collaboration of an ontology-based and autoregressive model-based chatbot is significantly effective.

Keywords: ontology; chatbots; answer-question systems; domain knowledge base; deep learning; autoregressive model.

Reference to this paper should be made as follows: Nguyen, T.T.S., Ho, D.H.T., Nguyen, N.T.A. and Do, P.M.T. (2024) 'A collaboration of an ontology and an autoregressive model to build an efficient chatbot model', *Int. J. Intelligent Information and Database Systems*, Vol. 16, No. 3, pp.241–257.

Biographical notes: Thi Thanh Sang Nguyen received her PhD degree in Software Engineering from the University of Technology, Sydney (UTS) in 2013. Her PhD thesis is about Semantic-enhanced Web-page Recommender Systems. She was supervised by Dr. Helen Lu and Prof. Jie Lu. She received her Masters degree in Computer Engineering from the University of Technology (VNU-HCMC) in 2006. She is working as a lecturer at the School of Computer Science and Engineering at the International University (VNU-HCMC) from 2015. She has more than 20 published research papers in the field of data mining. Her research interests include web mining, semantic web, knowledge discovery, deep learning and business intelligence.

Dang Huu Trong Ho is a Master student at the School of Computer Science and Engineering at the International University (VNU-HCMC).

Ngoc Tram Anh Nguyen received her Master degree in Information Technology Management at the School of Computer Science and Engineering at the International University (VNU-HCMC) in 2022.

Pham Minh Thu Do received her Masters degree in Information Technology Management at the School of Computer Science and Engineering at the International University (VNU-HCMC) in 2021.

1 Introduction

A chatbot is now not strange in our life, it becomes a friend, a consultant, or an assistant. In other words, a chatbot is able to understand and communicate with people and perform specific tasks. In natural language processing, it is used in applications that offer automatic verbal interactions. For example, a chatbot (Wu et al., 2020) has been developed alongside an E-learning platform in order to answer questions relevant to course materials, and chitchat as well. This makes online classes more interesting, especially, nowadays, online courses/classes are very popular. Depending on different architectures, these smart entities are capable of communicating in many ways, whether to provide instructions, answers to questions, or to entertain users.

According to Adamopoulou and Moussiades (2020), there are many types of chatbots, classified based on: the knowledge domain, the service provided, the goals, the input processing and response generation method, the human-aid, and the build method. In terms of the knowledge domain, this study concerns closed domain chatbots, focusing on university admissions and education. Besides, the input processing and response generation methods which are considered in this study are retrieval-based model and generative model. In the retrieval-based model, a domain ontology is able to be built for information retrieval since it is a powerful expression tool (Antoniou and Harmelen, 2008). Some ontology-based chatbots have been built to support answering automatically questions in specific domains, e.g., shopping in e-commerce (Hallili, 2014), drug information consultant in medical (Avila et al. 2019), and educational and professional orientation in education (Zahour et al., 2020). Theses chatbots help human much nowadays because of an overload of consultative jobs in many fields or being impracticable to access to a person in charge. In the generative model which is more human-like, some usable methods are machine learning or deep learning, e.g., recurrent neural network (RNN) (Cho et al., 2014), long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), Qanats (Yu et al., 2018), sequence-to-sequence (Seq2seq) (Jurafsky and Martin, 2020), hierarchical recurrent encoder – decoder (HRED) (Serban et al., 2016), SPHRED (Shen et al., 2017), let (Yang et al., 2020), etc. For example, by improving the QANet model to be combined with the retrieval-based model, a hybrid model K-12 e-learning assistant chatbot (Wu et al., 2020) was built and better than a teacher counselling service.

1.1 Problem statement

As we can see, the generative model is more interesting and effective for open domain chatbots, but it requires a large amount of well-prepared question-answer (QA) data for training the model. In many cases, this kind of data is not available at the beginning, so domain data is necessary to be collected to construct a knowledge base for a chatbot model, e.g., an ontology-based chatbot. This study proposes a new chatbot model for admissions advising at the International University (IU) belonging to the Vietnam National University – Hochiminh City (VNU-HCMC). A chatbot of advising admissions and education is very necessary nowadays in order to decrease the workload of academic advisors, but it has not been built yet in Vietnam and the QA data of Vietnam university admissions and education is not available or not enough. Therefore, a domain ontology is constructed given the information of university admissions of the university. Based on the domain ontology, we can make response reasoner in the specified domain. To enhance the ability of answering un-structured questions or ones having not enough keywords to retrieve information from the ontology, an autoregressive model is taken into account. And to train the autoregressive model, a set of QA pairs generated from the domain ontology are able to be used. However, training the Vietnamese QA of university admissions and education has not been done before, hence it is very challenging to do this work.

1.2 Objectives

The study focuses on building an innovative ontology-based chatbot with the support of autoregressive model which can overcome missing responses unable to be performed by the ontology-based model. Firstly, the proposed ontology model in the chatbot is designed newly and generally in order to be applied to similar application domains of advising admissions at different universities in Vietnam. Secondly, the autoregressive generative models which can understand natural languages, such as, XLNET, are developed in a new pipeline for handling Vietnamese language and then integrated into the ontology-based chatbot model. The combination of both the models can provide more correct answers than using one model. In this study, the admissions data of the IU (VNU-HCMC) is used for populating the proposed ontology and making training data for the autoregressive generative models. Moreover, the proposed chatbot can be extended by adding more data into the ontology to enrich the training data of the autoregressive model so that the performance of the chatbot will be able to be improved significantly.

The following sections will present related work (Section 2), research methodology (Section 3), experimental results with evaluation (Section 4) and conclusions (Section 5).

2 Related work

As mentioned in the introduction section, this section presents related techniques and the concerning chatbot models which are based on autoregressive models and ontology.

2.1 Ontology

According to Antoniou and Harmelen (2008), ontology is a knowledge representation technology, in that, concepts and relations between them are defined in one or different domains. It is efficient to express the semantic information of a knowledge base. That is why ontology is used to represent semantic knowledge bases for automatic inference or information retrieval in a specific domain. OWL (Web ontology language, http://www.w3.org/TR/owl-features/) is a main Web ontology language which satisfies the requirements of building a domain ontology, including a well-defined syntax, a well-defined semantics, efficient reasoning support, sufficient expressive power, and convenience. Therefore, an ontology can be used efficiently in a search engine of a chatbot.

2.2 Autoregressive models

Autoregressive models are known as a deep generative model or a feed-forward model which can predict future values given past values. This prediction is based on computing the maximum likelihood of an event in a sequence given the previous events. In detail, the observations were given as a sequence ($x^{(1)}$, …, $x^{(T)}$), the likelihood is decomposed into a product of conditional distributions:

$$p\left(x^{(1)},\ldots,x^{(T)}\right)=\prod_{t=1}^{T}p\left(x^{(t)}\left|x^{(1)},\ldots,x^{(t-1)}\right.\right).$$

Autoregressive models might be another good choice of building a knowledge base of a chatbot but need a good training dataset.

2.3 Existing chatbot models

2.3.1 Autoregressive model-based

As known, autoregressive models are modern deep learning models for producing more human-like responses in chatbots but are very challenging to build. We cannot deny that preparing knowledge for chatbot is the most vital task but also a time-consuming one. Arsovski et al. (2019) proposed a method of building a chatbot that could reduce the time spending on acquiring knowledge. Their attempt included two phases that are extracting conversational knowledge (phase 1) and building a chatbot using neural networks (phase 2). The main contribution of this study was they found out the saturating pattern of the number of the input questions then clustered the answers using K-means. In phase 2, the chatbot was built using the extracted information from the previous phase. The proposed model was a sequence-to-sequence LSTM Neural Network framework. The result of this approach was quite unremarkable: the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) score was greater than 0.5 for 75% of the entries. In addition, they concluded that the suggested novel approach might be utilised to force conversational knowledge extraction for any type of chatbots.

In 2020, Dhyani and Kumar has created an open-domain Chatbot that may subsequently be subjected to any specific domain (if necessary). This domain-expanding can be accomplished by changing the dataset, which trains a model with particular domain knowledge. While, the existing methods face with the limitation of vanilla RNN

that is being defeated by long sequences. Another challenge of RNN is its unidirectional, which cannot capture the context of a word in a sentence as good as a human does in a natural language. The approach of using BRNN (Bi-Directional Neuron Network) along with Attention mechanism was introduced to overcome the listed obstacles.

In Vietnamese language, no publication had been yet studied on evaluating what users say in chatbots in depth. Therefore, Tran and Luong (2020) focused on analysing and constructing an intelligent module that will allow the bot to better interpret user utterances. It proposed the framework of analysing Vietnamese utterances including two key modules: an intent parser and a context extractor. Regarding the Intent Parser, they proposed the methods of combining LSTMs and CNNs. Two LSTMs are trained on user utterances to increase model performance. The first focuses on a left-to-right utterance, whereas the second focuses on reversed replica of the considering utterance. By default, the forward and backward outputs should be concatenated before being sent to the next layer. Finally, the prediction is obtained by applying an activation function to the concatenation vector. For each feature map, the goal of using CNN is to capture the most significant features with the highest value. These feature vectors are then fed into the final activation layer to classify the utterance. For the context extraction, they suggested the approach of using nonlinear neural networks, such as LSTMs and CNNs, to automatically learn features for conditional random fields (CRFs) without having to construct them by hand. Aside from a tiny amount of supervised training data and unlabeled data, these models have no language-specific resources. The results showed that, in general, utilising neural networks might increase the performance of a system over a traditional one. Overall, they used CNNs to get the best F-measure of 82.32 percent in identifying intentions.

Recently, a well-known model, i.e., XLNet (Yang et al., 2020), outperforms BERT (Devlin et al., 2018) by proposing a permutation language model which captures bidirectional contexts to learn all combination of inputs and predict words arbitrarily. The model is trained with two-stream self-attention involving content and query streams. Moreover, the authors extended the ideas from transformer-XL model (Dai et al., 2019), i.e., the relative positional encoding scheme and the segment recurrence mechanism, into pretraining the model in order to predict a part of the input sequence and be able to extract dependent pairs.

2.3.2 Ontology-based

The weaknesses of the above models are dependence on training data of QA pairs. In cases the training data is not available or not good, it is necessary to build a knowledge base for a chatbot in a specific domain. Besides, learning from training data often lacks semantic information or expert knowledge. This leads to some difficulties in recognising context objects or application domains. Ontology is a solution for understanding what utterances are about. That is the reason ontology-based chatbot models were born. It is also driven domain knowledge so that it can create domain-driven conversations. Ontologies are used to store the domain knowledge and navigate through domain (Altinok, 2018). Therefore, this ontology-based knowledge base can provide information for answer generation in dialogs. The benefit of the ontology-based approach is to “keep conversation memory explicitly throughout the conversation”. Because of the benefits of ontology, ontology has been adopted into closed domain chatbots and provides very specific answers given questions.

For examples, an e-learning bot built by Clarizia et al. (2018) allows dealing with students' questions of subjects in lectures. Its knowledge base is an ontology containing 'users' and 'learning objects' which is a collection of content items, practice items, and assessment items. It plays a role as an education support system for students. Experimental results show that the chatbot can furnish about 71% correct suggestions. In another application domain, such as medical, an ontology of drugs and their relevant information has been constructed for MediBot (Avila et al., 2019), which is Portuguese Speakers Drug. The knowledge base of the ontology is a combination of many data sources and expert knowledge. The bot is responsible for converting natural language to SPARQL query, processing the query, and sending a response to users. Furthermore, Adhikary et al. (2022) also proposed an Ontology-Based healthcare hierarchy integrated into a chatbot for searching medical terminologies. It is useful for sharing medical knowledge.

As seen, the both generative and ontology-based models have different benefits. The generative model is more flexible and able to learn from different knowledge domains but needs a good training dataset. The ontology-based model is limited to a specific knowledge domain, but able to be built from the real-world data sources with domain experts' support. For instance, an ontology-based chatbot could enhance experiential learning through providing questions/answers pairs (dialogues) to train a deep neural network model in a cultural heritage scenario (Casillo et al., 2022). This chatbot can deliver the most effective answer given a question. However, generating answers by the combination of ontology and generative models has not been interested much, and not been implemented for university admissions advising. Therefore, this study takes into account the both models to handle the case study of admissions advising at the IU.

3 Research methodology

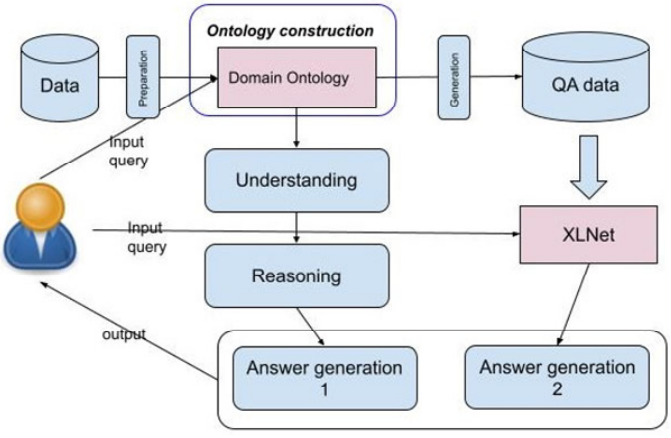
The proposed chatbot framework, namely OntoGen Chatbot, consists of two main process units: Ontology construction and Autoregressive modelling (Figure 1). In the ontology construction unit, an ontology of university admissions information is constructed for understanding input queries in natural language and reasoning relevant things in order to generate most suitable answers. On the other hand, to have additional responses for questions the ontology model did not find any answers, the autoregressive modelling unit, i.e., XLNet model, is employed. To train the autoregressive model, the domain ontology is used to generate question-and-answer pairs or conversations to be learned by the XLNet model. Basing on the trained autoregressive model, answers can be generated given the input queries. As a result, the generated answers are combined to output the final one. The following subsections will give more details of the proposed framework.

3.1 Ontology construction

As known, ontology technology has the expressive power in semantic knowledge representation. Based on that, this study takes into account building an ontology of a knowledge domain of university admissions as a case study. Particularly, the raw data is collected from the admissions website of the IU (https://tuyensinh.hcmiu.edu.vn/). It is cleaned to remove meaningless words, then analysed to construct ontology concepts and

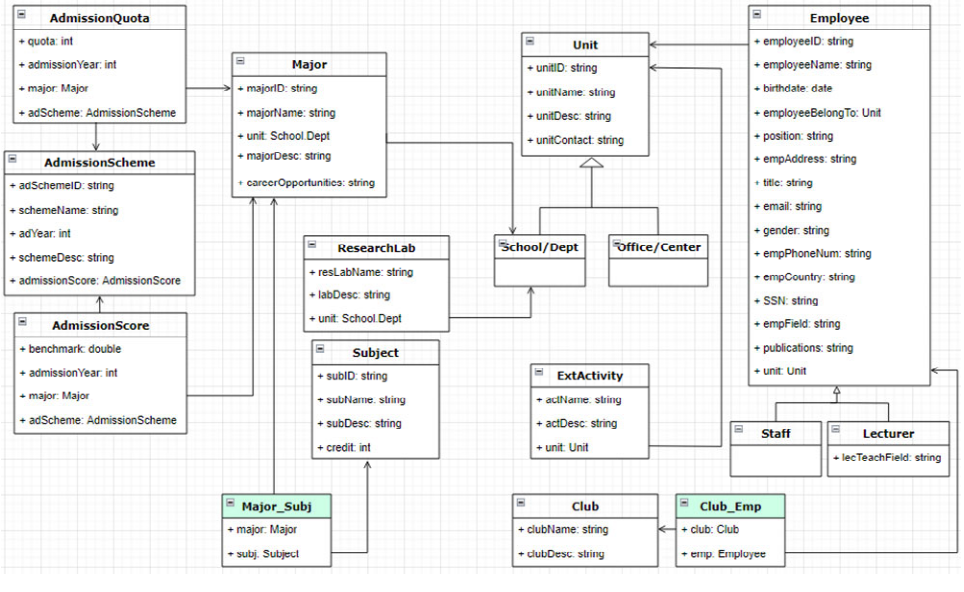
relationships among concepts in the preparation step before modelling a domain ontology.

Figure 1 The proposed chatbot framework (see online version for colours)



To model the domain ontology, domain entities are first identified, and then data and object properties are added. Figure 2 depicts the ontology model of the IU, including Major, Unit, Employee, etc. This study focuses on the enrollment and academic advising, so the information of programmes is considered along with subjects.

Figure 2 Ontology model of the IU (see online version for colours)



The following is the definition of the domain ontology model, namely IUOnto.

Definition 1. (Domain ontology model of the IU)

A domain ontology structure of IU is defined as a four-tuple: $O := < C, R_{SUB}, P, A >$, where C represents classes and entities in the IU domain, R_{SUB} describes SubClass-Of relationships, P represents properties defined in the ontology, and A represents axioms, such as, an instantiation axiom assigning an instance to a class, an assertion axiom assigning two instances by means of a property, a domain axiom for a property and a class, and a range axiom for a property value and an instance. In detail, C and P are further divided into sets:

$C = C \cup I$ comprises a set of domain classes (concepts) C, and a set of specific domain instances (of the concepts) I. The identified domain entities of university admissions are mainly majors, subjects, admissions schemes, admission quotas and scores of majors, and relevant ones, e.g., units, employees, research labs, clubs, and other activities. As shown in Figure 2, there are 16 classes categorised into three groups: C_1 including classes without parents, e.g., school/dept, employee, major, subject; C_2 including association classes, e.g., Major_Subj and Club_Emp are association classes between the two connected classes Major – Subject and Club – Employee, respectively; C_3 including classes having sub-classes, e.g., Unit. R_{SUB} comprises a set of the SubClass-Of relationships: School/Dept and Office/Center are the subclasses of Unit class, Staff and Lecturer are the subclasses of the Employee class.

$P = R \cup A$ comprises a set R of object properties in classes (C), and a set A of data properties. For each major, there are admission scores and quotas with respect to admission schemes. Each unit manages a number of majors and employees as well as academic activities. Particularly, in IUOnto, the Major and Unit object properties in the AdmissionQuota and Employee classes, respectively, refer to the corresponding Major and Unit classes, as Figure 2. The adScheme is an AdmissionScheme object property of AdmissionQuota and AdmissionScore classes. The object properties often have their own inverses to facilitate querying information from both sides. Each class has some data properties, e.g., Lecturer has the lecTeachField attribute with *string* type and the attributes inherited from the Employee class.

The defined concepts in IUOnto are generic so that they can represent the knowledge of university admissions advising in Vietnam in general speaking and in the IU in specific speaking. Based on IUOnto, the IU ontology[1] is built using Protégé tool, as shown in Figure 3.

To quickly seek instances in the IU ontology, some keywords are added into each instance. Therefore, a multivalued keywordName property is added into the Thing class. Moreover, some keywords are commonly used in many instances, hence we have the Keyword class. Besides, keywords categorised in groups are presented by the KeywordByClass class.

From this ontology, we can construct a full ontology-based knowledge base by populating the data collected from the IU into the built ontology. For instances, Figure 4 shows some instances of AdmissionScore and Subject classes. For classes that do not have a Name property, e.g., AdmissionScore, their IDs must be meaningful names.

Figure 3 The admissions ontology of the IU (see online version for colours)

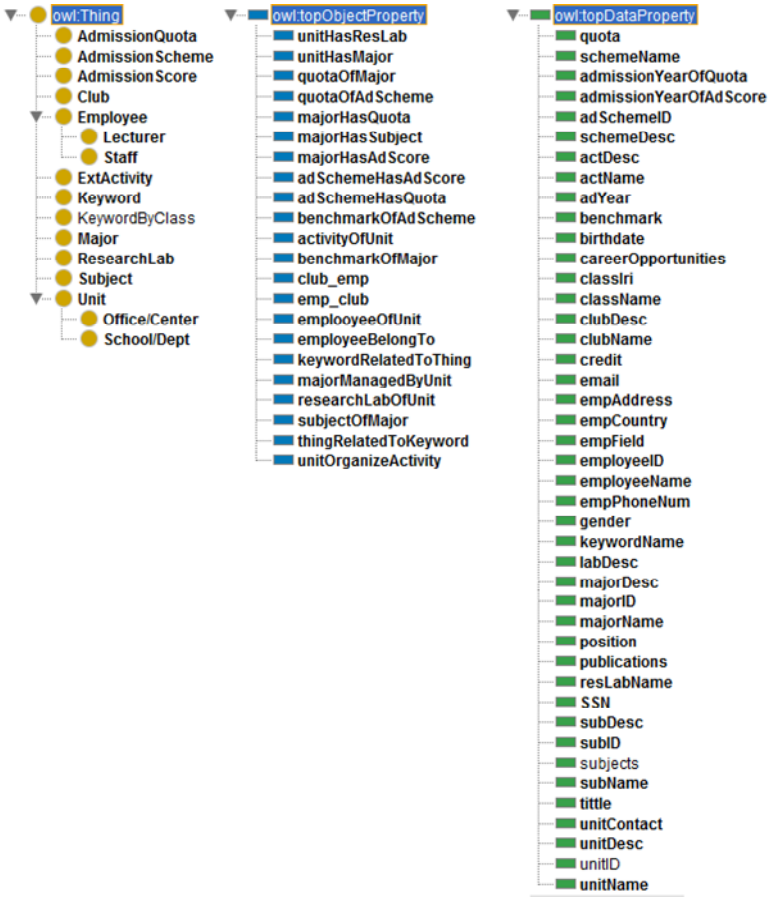


Figure 4 The partial instances of AdmissionScore and Subject classes (see online version for colours)

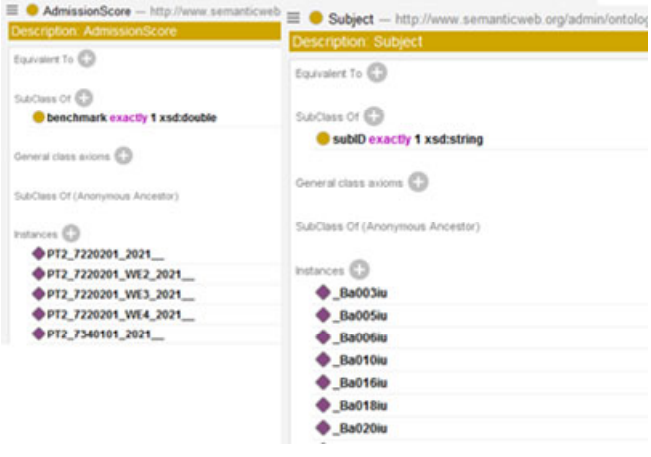
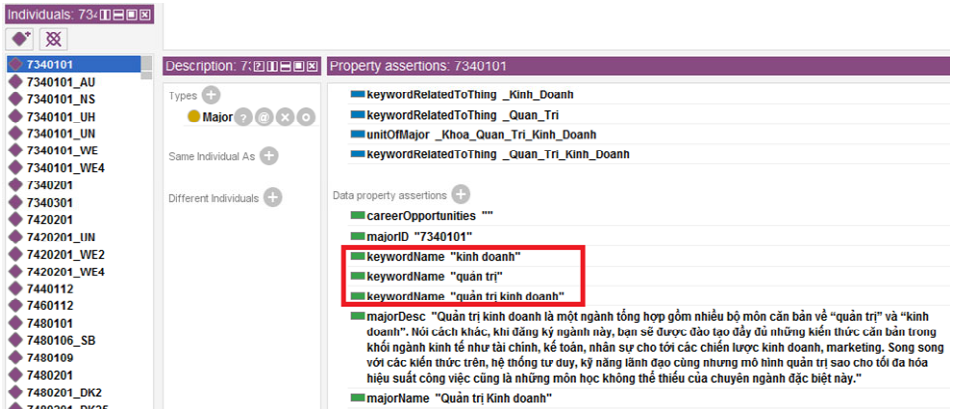


Figure 5 shows some keywords added into the instances of Major class. It is noted that data values in the ontology are in Vietnamese, since the IU admissions data is in Vietnamese. In Figure 5, the 'Quản trị Kinh Doanh' (*Business Administration*) Major has some keywords 'kinh doanh' (*business*), "quản trị" (*administration*), and 'quản trị kinh doanh' (*business administration*).

Figure 5 The partial instances of the major class and relevant keywords (see online version for colours)



Reasoning and answer generation

Based on the above ontology, the machine can reason information of instances so that it can understand the meaning of a keyword if that one is found in the list of keywords or instance names/IDs. By ontology reasoning, the OntoGen chatbot can generate answers to users' questions. Algorithm 1 presents how to generate answers given a question. Keywords from questions are extracted by using the VNcoreNLP library (Vu et al., 2018). The output answer is the information or description of the returned instance.

Algorithm 1 Ontology-based answer generation

```
Input: A question from a user
Output: A set of relevant answers
Process:
Function ask(question)
{
Extract keywords in the given question by the VNCoreNLP library
List<string> userKeywords = extractKeywords(question);
Let relavantThings = {}; // a set of instances found by keywords
(may be duplicated)
      For each userKeyword in the user's question
            For each owlKeyword in the ontology
                  If(userKeyword matches owlKeyword)
                  {
```

```
                    Get all instances {I} by owlKeyword;
                    relavantThings.add(I);
                }
    Group relevantThings by IRI to count the number of matches
of each instance
    Sort the grouping result in descending order of the
matching numbers;
    Return relevantThings with the most matches;
}
```

3.2 Autoregressive modelling

As mentioned above, to generate more answers to questions whose keywords are not found from the list of keywords in the ontology, an autoregressive model is employed. However, it is necessary to obtain a dataset of QA pairs in the application domain for training the model. Hence, the IU ontology (IUOnto) can be used to generate QA data.

QA data generation

Since we focus on the questions of admissions, all instances of the Major and Unit classes are taken into account. The templates of input questions are: “May I ask about <major name>|<unit name>?”, “What is the admissions quota of <major name>?”, “What is the admissions score of <major name>”. By reasoning the IUOnto, responses are generated. The generated QA pairs are grouped into majors or units. Each of majors or units is considered as the topic of a conversation including many QA pairs. Given the IU ontology data, 167 QA pairs are generated. The partial QA data is as follows (the QA content translated into Vietnamese is put in the parentheses):

Q: Cho tôi hỏi về Công nghệ Thực phẩm?

(*May I ask about Food Technology?*)

A: Ngành Công nghệ Thực phẩm (Mã ngành: 7540101) Chương trình: Trong nước, Chương trình đào tạo ngành CNTP tại Trường Đai học Quốc tế được thiết kế nhằm đào tạo các kỹ sư CNTP, có những nét chính như sau: Cung cấp kiến thức rộng, cân bằng về các khoa học thực phẩm và kỹ thuật thực phẩm, cung cấp kiến thức cơ bản về quản lý thực phẩm (bao gồm cả quản lý sản xuất, chất lượng, vệ sinh an toàn thực phẩm và vệ sinh môi trường), chú trọng đến kỹ năng thực hành và khả năng tiếp cận với thiết bị công nghệ cao. Giảng dạy hoàn toàn bằng tiếng Anh (kỹ sư sau khi tốt nghiệp sẽ đạt trình độ tiếng Anh TOEF pBT 550 hoặc tương đương (IELTS 6.0, TOEFL iBT 60).

(*Food Technology (Code: 7540101) Program: Domestic, The food technology program at the International University is designed to create food technology engineers, with the following main features: Providing the broad and balanced knowledge of food sciences and food engineering, providing basic knowledge of food management (including production, quality, food safety and hygiene, and environment hygiene management), emphasis on practical skills and access to high-tech equipment. Teaching is entirely in English (engineers graduate with TOEF pBT 550 or equivalent (equivalent IELTS 6.0, TOEFL iBT 60).*)

XLNet (Yang et al., 2020)

The XLNet model is the autoregressive model used in the OntoGen chatbot. Since the studied training data is in Vietnamese, it is necessary to find a suitable tokenizer tool. HuggingFace[2], an open-source platform in machine learning, offers Transformers providing thousands of pretrained models for Natural Language Processing (NLP). In that, XLNet was pretrained in English, while BERT (Devlin et al., 2018) was pretrained in Multilanguage (including Vietnamese). According to Yang et al. (2020), the XLNet model outperforms the limitations of the BERT model. Therefore, this study designs a new pipeline, in which, the tokenisation task is performed by Multilingual BertTokenizer for handling Vietnamese language, and the language modelling task is performed by XLNetModel. This pipeline is called Bert-XLNet and is fine-tuned for the new data of admissions advising.

The implemented model includes four steps:

1 preprocessing the training data

2 fine-tuning the model

3 training the model

4 evaluating answers.

In the preprocessing step (1), text tokenisation, context truncation, mapping the positions of questions, answers and context are performed. Particularly, the QA data is reformatted to be fed into the model. Moreover, it is necessary to provide a context of each QA pair for training. Utilising IUOnto, the context of a question can be identified. Particularly, the major/unit related to the keywords in the question are able to be returned as topics. The QA pairs are grouped by topics. For each group, answers are merged to form a context. Because the context length of each question is limited to 512 characters, they are splitted into shorter context. Each context is then combined with the corresponding QA pair. As a result, there are more than 2000 tuples of {question, answer, context} made for training the model.

In the fine-tuning step (2), the training arguments of the model are optimised. It is necessary to fine-tune the autoregressive model for a number of epochs until obtaining parameters most suitable for the training data. In the training step (3), training loss and validation loss are minimised to obtain the good model. In the evaluation step (4), prediction results are the answers having highest scores which is computed based on the span-start scores (start-logits) and span-end scores (end-logits). To predict an answer for each question, we take the index of the maximum start_logits as a start position, and the index of the maximum end_logits as an end position. However, there are several outlier cases, such as, the start position could be greater than the end position, or the start position points to a span of text in the question instead of the answer. To deal with them, we would look for the second-best prediction (with the second-best score) to see whether it gives a possible answer or not. In case the second-best case does not satisfy the prediction, we continue the process until finding out the possible answer. To classify the answers, the score achieved by summing the start and end logits will be used. We do not try to arrange all of the potential responses; instead, we use a hyper-parameter to limit themselves. The best indices in the start and end logits will be chosen, and all the predicted answers will be gathered. They will be ranked by scores and the top one will be kept after they are all legitimate.

4 Experimental results and evaluation

In this study, four experimental cases of are carried out to examine the chatbot models in the proposed framework. Some questions at different levels of difficulty are prepared for the experiments. We will validate the proposed framework in the following cases.

- Case 1: The ontology-based model is performed. The responses are generated from the IUOnto only.
- Case 2: The XLNet-based model is examined. The responses are generated from the XLNet model using XLNetTokenizer. This case is the original pipeline of XLNet model implementation.
- Case 3: The Bert-XLNet-based model is examined. The responses are generated from the Bert-XLNet pipeline using BertTokenizer.
- Case 4: A combination of Case 1 and 3. The responses are generated from the best-chosen results. This is the case of the proposed framework.

4.1 Training and testing datasets

Since the OntoGen chatbot is used for admissions advising at IU, the used dataset is collected from the information of the IU admissions website https://tuyensinh.hcmiu.edu.vn/. Given the IU data, the IU ontology is constructed. The programmes of 21 majors at IU along with their subjects are added into the ontology. The admission score and quota of each major are also modified. There are six admissions schemes, 12 schools/departments and 17 offices/centers appended into the ontology. As mentioned above, 167 QA pairs are generated for training the XLNet models. Depend on the used Tokenizer, the number of tuples of {question, answer, context} made in the preprocessing step is different. When the XLNetTokenizer is used, there are 2542 tuples of {question, answer, context} are made for training and 288 tuples for validation. When the BertTokenizer is used, there are 3053 tuples of {question, answer, context} are made for training and 347 tuples for validation. The set parameters of the XLNet model are max_seq_length=512, train_batch_size=8, predict_batch_size=32, eval_batch_size=8, seed=42, learning_rate=2e-5, optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08, iterations=100, save_steps=100, train_steps=1000, warmup_steps=100. Table 1 shows the training results of XLNet and Bert-XLNet-based models in the Huggingface platform.

Table 1 The training results of XLNet and Bert-XLNet in the Huggingface platform

Training loss		*Epoch*	*Step*	*Validation loss*	
XLNet	*Bert-XLNet*			*XLNet*	*Bert-XLNet*
No log	No log	1.0	382	0.500534	0.4695
0.822200	0.5633	2.0	764	0.448776	0.3361
0.822200	0.3533	3.0	1146	0.496503	0.3489

As seen, the training and validation loss of Bert-XLNet-based model are less than the ones of XLNet-based model. For testing the proposed models, a set of 16 questions related to the domain information in the learning sources are suggested as listed in

Table 2. These questions are frequent and typical when advising admissions at IU. These questions are classified at three level of difficulty:

1 easy (directly related to the application domain)

2 medium (relatively related to the application domain)

3 difficult (not relevant much to the domain).

Table 2 Experimental questions at different difficulty levels

No.	*Questions in Vietnamese*	*Level*
1	Các phương thức tuyển sinh năm 2021? *(Admissions schemes in 2021)*	1
2	Giới thiệu trung tâm dịch vụ công nghệ thông tin? *(Introduction to the center for information services)*	1
3	Làm thế nào lấy lại tài khoản email? *(How to get my email account back?)*	2
4	Cho tôi hỏi học phí trung bình là bao nhiêu? *(What is the average tuition fee?)*	3
5	Khoa quản trị kinh doanh đào tạo những ngành nào? *(What programs does the School of Business Administration offer?)*	1
6	Giới thiệu ngành quản trị kinh doanh? *(Introduction to business administration program)*	1
7	Chỉ tiêu tuyển sinh ngành quản trị kinh doanh. *(The admissions quota of the business administration major)*	1
8	Điểm chuẩn ngành quản trị kinh doanh. *(Admissions score of the business administration major.)*	1
9	Chỉ tiêu ngành công nghệ thông tin? *(The admissions quota of the information technology major)*	1
10	Điểm chuẩn ngành công nghệ thông tin năm 2021? *(Admissions score of the information technology major in 2021)*	1
11	Ngành ngôn ngữ anh ra trường làm gì? *(After graduating from an English linguistics program, what will a graduate be able to do?)*	2
12	Các ngành liên kết của khoa công nghệ thông tin? *(What are twinning programs at the school of computer science and engineering?)*	1
13	Hãy giới thiệu về thư viện của trường? *(Introduction to the library at the university)*	3
14	Phòng công tác sinh viên có chức năng gì? *(What are the functions of Students Services Office)*	1
15	Ngành công nghệ thông tin có những chuyên ngành nào? *(What are the majors in the Information Technology program?)*	2
16	Trường có những hoạt động ngoại khóa nào dành cho sinh viên? *(What extracurricular activities does the university have for students?)*	3

Table 3 The evaluation of satisfactory (1), correctness (2) and usefulness (3) in the experimental cases

Q. no.	*Case 1*			*Case 2*			*Case 3*			*Case 4*		
	(1)	*(2)*	*(3)*	*(1)*	*(2)*	*(3)*	*(1)*	*(2)*	*(3)*	*(1)*	*(2)*	*(3)*
1	*5*	*5*	*5*	0	0	0	0	0	0	5	5	5
2	*5*	*5*	*5*	4	3	4	0	0	0	5	5	5
3	*4*	*3*	*5*	4	3	4	0	0	0	4	3	5
4	0	0	0	0	0	0	*4*	*3*	*4*	*4*	*3*	*4*
5	*3*	*2*	*4*	0	0	0	0	0	0	3	2	4
6	*4*	*4*	*5*	3	1	3	3	1	3	4	4	5
7	*5*	*5*	*5*	2	0	2	4	1	4	5	5	5
8	*5*	*5*	*5*	0	0	0	0	0	0	5	5	5
9	*5*	*5*	*5*	0	0	0	5	4	5	5	5	5
10	*5*	*5*	*5*	0	0	0	0	0	0	5	5	5
11	*3*	*3*	*4*	0	0	0	0	0	0	3	3	4
12	*5*	*5*	*5*	3	2	3	3	2	3	5	5	5
13	0	0	0	*3*	*2*	*3*	*3*	*2*	*3*	*3*	*2*	*3*
14	*5*	*5*	*5*	0	0	0	0	0	0	5	5	5
15	*5*	*4*	*5*	2	1	2	2	1	2	5	4	5
16	*0*	*0*	*2*	0	0	0	0	0	0	0	0	2
Avg.	*3.6875*	*3.5*	*4.0625*	*1.3125*	*0.75*	*1.3125*	*1.5*	*0.875*	*1.5*	*4.125*	*3.8125*	*4.5*

4.2 Experimental results

According to Majid et al. (2021), the evaluation criteria of responses are able to be the rates of satisfactory, correctness and usefulness represented by scores from 1 to 5. The responses from the chatbot are remarked by domain experts. The highest score (5) means the response is completely satisfied, correct or useful. The medium score (3) means somehow the response may give some acceptable, meaningful, useful information. The lowest score (0) means the response is completely not relevant or useful. Table 3 presents the evaluation scores of responses to the above questions in the four experiment cases.

4.3 Evaluation and discussion

As shown in Table 3, the ontology-based model (Case 1) can provide most of the responses satisfied high, except for difficult questions (4, 13, 16). Moreover, the responses are meaningful and useful. While, the XLNet model do not give full answers, only few meaningful keywords. As we can see, Case 3 is better to generate considerable responses than Case 2. Especially, the Bert-XLNet-based model can make the responses on Question 4 and 13 with high satisfactory levels (score 4 and 3, respectively), which helps to support answering the difficulty questions in Case 1. Case 4 which is a combination of the ontology and Bert-XLNet-based models can get benefits or best responses from the models.

In the context of autoregressive model, it is known that frameworks are very hard to train. Because there is not much data for training and the training answers are quite long, it is very challenging for the XLNet-based models to train this kind of data. They are not

able to provide long answers. Moreover, they are not pretrained much in Vietnamese corpuses, so there are many limitations in these models. Therefore, the autoregressive models can be used as support tools to generate answers for some questions which are not responded effectively by the ontology-based model in this study.

In the context of the ontology model, the IUOnto contains essential concepts for university admissions, more variety than the ontology of educational program counselling built in Majid et al. (2021) which has only the concepts of City/Capital, Programe, DeptPrograme and Type. Moreover, this ontology can be extended with more data so that more responses can be generated. The QA pairs generated from the IUOnto for training the autoregressive models were checked and all of them are correct. It proves that the IUOnto is useful and reasonable.

5 Conclusions

The XLNet-based models can generate responses based on the context assigned to QA pairs. These models will make better answers when the context of a conversation is more specific. However, they depend much on the training data and require a large of training data. Thanks to the IUOnto, the training data can be prepared at the beginning, and the context of questions can be identified to facilitate the autoregressive models. Besides, the ontology-based model can achieve high performance in answering questions in the learned application domain. The combination of the autoregressive model and the ontology-based model can enhance the performance of the OntoGen chatbot, as Case 4. The experimental results have shown the proposed chatbot framework is promising.

In the future, the chatbot will be improved when more data is trained, and the QA pairs are refined for training the autoregressive models.

References

Adamopoulou, E. and Moussiades, L. (2020) *An Overview of Chatbot Technology*, Springer International Publishing, Cham.

Adhikary, P.K., Manna, R., Laskar, S.R. and Pakray, P. (2022) *Ontology-Based Healthcare Hierarchy towards Chatbot*, Springer International Publishing, Cham.

Altinok, D. (2018) 'An ontology-based dialogue management system for banking and finance dialogue systems', *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, European Language Resources Association (ELRA).

Antoniou, G. and Harmelen, F.V. (2008) *A Semantic Web Primer*, MIT Press, USA.

Arsovski, S., Osipyan, H., Oladele, M.I. and Cheok, A.D. (2019) 'Automatic knowledge extraction of any Chatbot from conversation', *Expert Systems with Applications*, Vol. 137, pp.343–348.

Avila, C.V.S., Calixto, A.B., Rolim, T.V., Franco, W., Venceslau, A.D.P., Vidal, V.M.P., Pequeno, V.M. and Moura, F.F.D. (2019) MediBot: an ontology based chatbot for Portuguese speakers drug's users', *Proceedings of the 21st International Conference on Enterprise Information Systems* – Volume 1: ICEIS, Heraklion, Crete, Greece, pp.25–36.

Casillo, M., De Santo, M., Mosca, R. and Santaniello, D. (2022) 'An ontology-based chatbot to enhance experiential learning in a cultural heritage scenario', *Frontiers in Artificial Intelligence*, 25 April, Vol. 5, pp.1–18.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*, Association for Computational Linguistics, Doha, Qatar.

Clarizia, F., Colace, F., Lombardi, M., Pascale, F. and Santaniello, D. (2018) *Chatbot: An Education Support System for Student*, Springer International Publishing, Cham.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V. and Salakhutdinov, R. (2019) *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*, arXiv:1901.02860v3 [cs.LG].

Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, CoRR abs/1810.04805.

Dhyani, M. and Kumar, R. (2020) 'An intelligent Chatbot using deep learning with Bidirectional RNN and attention model', *Materials Today: Proceedings*.

Hallili, A. (2014) 'Toward an ontology-based chatbot endowed with Natural Language processing and generation', *26th European Summer School in Logic, Language & Information*, Tübingen, Germany.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Comput.*, Vol. 9, No. 8, pp.1735–1780.

Jurafsky, D. and Martin, J.H. (2020) 'Machine translation and encoder-decoder models', in Jurafsky, D. and Martin, J.H. (Eds.): *Speech and Language Processing*.

Majid, M., Hayat, M-F., Khan, F-Z., Ahmad, M., Jhanjhi, N., Bhuiyan, M-A-S., Masud, M. and AlZain, M-A. (2021) 'Ontology-based system for educational program counseling', *Intelligent Automation & Soft Computing*, Vol. 30, No. 1, pp.373–386.

Papineni, K., Roukos, S., Ward, T. and Zhu, W-J. (2002) 'BLEU: a method for automatic evaluation of machine translation', *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, pp.311–318.

Serban, I.V., Sordoni, A., Bengio, Y., Courville, A. and Pineau, J. (2016) 'Building end-to-end dialogue systems using generative hierarchical neural network models', *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona.

Shen, X., Su, H., Li, Y., Li, W., Niu, S., Zhao, Y., Aizawa, A. and Long, G. (2017) *A Conditional Variational Framework for Dialog Generation*, CoRR abs/1705.00316.

Tran, O.T. and Luong, T.C. (2020) 'Understanding what the users say in chatbots: a case study for the Vietnamese language', *Engineering Applications of Artificial Intelligence*, Vol. 87, p.103322.

Vu, T., Nguyen, D.Q., Dras, M. and Johnson, M. (2018) 'VnCoreNLP: a Vietnamese natural language processing toolkit', *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, New Orleans, Louisiana, pp.56–60.

Wu, E.H., Lin, C., Ou, Y., Liu, C., Wang, W. and Chao, C. (2020) 'Advantages and constraints of a hybrid model K-12 e-learning assistant chatbot', *IEEE Access*, Vol. 8, pp.77788–77801.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V. (2020) *XLNet: Generalized Autoregressive Pretraining for Language Understanding*, arXiv:1906.08237 [cs.CL].

Yu, A.W., Dohan, D., Luong, M-T., Zhao, R., Chen, K., Norouzi, M. and Le, Q.V. (2018) 'QANet: combining local convolution with global self-attention for reading comprehension', *Proc. ICLR*, Vancouver, BC, Canada.

Zahour, O., Benlahmar, E.H., Eddaoui, A., Ouchra, H. and Hourrane, O. (2020) 'A system for educational and vocational guidance in Morocco: Chatbot E-Orientation', *Procedia Computer Science*, Vol. 175, pp.554–559.

Notes

1 https://doi.org/10.6084/m9.figshare.22257136.v1.

2 https://huggingface.co.