IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 July 2023, accepted 15 August 2023, date of publication 22 August 2023, date of current version 28 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307635

RESEARCH ARTICLE

A Transformer-Based Educational Virtual Assistant Using Diacriticized Latin Script

KHANG NHUT LAM[1], LOC HUU NGUY[1], VAN LAM LE[1], AND JUGAL KALITA[2]
[1]Department of Information Technology, Can Tho University, Can Tho 94100, Vietnam
[2]Department of Computer Science, University of Colorado, Colorado Springs, CO 80918, USA

Corresponding author: Khang Nhut Lam (lnkhang@ctu.edu.vn)

ABSTRACT A virtual assistant or smart chatbot should be able to understand user questions and respond correctly and usefully, even if the questions are posed ungrammatically with misspellings and other errors. This paper describes the design and construction of a text-to-text virtual assistant in Vietnamese, a language that uses the Latin script with a liberal use of diacritics, for supporting students at a large university with over forty thousand students. The flexible virtual assistant consists of two integrated chatbots, both using Transformers: a) a closed-domain chatbot, trained on over thirty-five thousand factual question-answer pairs to engage in university-related conversation, and b) a second open-domain chatbot, trained on a large movie dialog dataset to engage in general conversation. The integrated virtual assistant classifies a question as either factual or general, and engages the appropriate chatbot to respond in a flexible, appropriate and natural manner. Although Vietnamese uses diacritics copiously, even educated users have a propensity to forgo the use of diacritics, and as a result, to facilitate smooth text-based communication, our design includes extensive pre-processing that uses learned Transformers to restore missing diacritics and correct misspellings. Our Transformer models outperform existing approaches for diacritic restoration and are better than several other methods for spelling correction in Vietnamese. In addition, the closed-domain chatbot performs better than other generative chatbots that have been developed to assist students in a university environment, irrespective of language and location.

INDEX TERMS Chatbot, diacritic restoration, educational chatbot, misspelling, transformer, virtual assistant, Vietnamese.

I. INTRODUCTION

Virtual assistants, or chatbots, have exploded in use to support business activities, answer questions related to citizen public services, and provide answers to users such as patients or students. This paper discusses building a smart chatbot using neural network approaches. One can construct a chatbot using retrieval-based or generation-based neural network methods. Chatbots built using the first method may give answers with correct grammar and spelling, whereas the second method may or may not. However, if users ask questions that are not in the training dataset, a chatbot built using the retrieval-based method cannot answer them while the other can [1].

Our goal is to construct a chatbot to support students in a university environment such that it can generate new responses based on inputs and does not depend on solution provided by third party vendors. Moreover, Caldarini et al. [1] claim that most existing chatbots in the domain of education are constructed using retrieval-based models, and as a result, cannot answer questions outside a set of pre-defined question-answer pairs. Therefore, constructing a chatbot using the generation-based approach is the most suitable choice for us. Although students can find information in many ways, such as perusing websites and asking friends, professors or advisors, many students do not want to talk to others or do not know where to find information (a conclusion based on a limited informal survey). As a proof-of-concept, we build a text-to-text virtual assistant, named VietBOT, for students in Vietnamese. The chatbot answers questions related to programs, staff, academic regulations, courses, semester examination schedules, and so on.

Raval [2] highlights the challenges facing chatbots, including understanding intents behind questions written in various writing styles, handling local vocabularies, synonyms, and

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

sentiment analysis. For many languages using Latin script such as Vietnamese, diacritics are extremely important; missing or typing incorrect diacritics may change the meaning of a word. However, it is normal for students to type questions with words without using diacritics, or using misspelled words. Therefore, we propose an approach to help the chatbot handle questions with these issues. In addition, our chatbot model trains faster than other existing models for constructing a chatbot to assist students in Vietnamese.

The contributions of this paper are enumerated below.

- A Transformer-based model for diacritic restoration that can be used for languages using Latin script with diacritics.
- A Transformer-based model for spelling correction that handles mistyped and misspelled words created by several text input approaches for a highly diacriticized language, viz., Vietnamese.
- An architecture for a generative method using Transformer for constructing a chatbot to assist students in a university environment, and Transformer-based models for handling questions with various error levels in user input.
- A method using the kNN algorithm to combine closed and open domain chatbots in one application to support students.
- Suggestion of methods to construct a bank of question-answer pairs for training a closed chatbot in a university domain.
- Two Vietnamese datasets that can be used to train chatbot models in both closed and open domains.

The remainder of this paper is organized as follows. Section II presents related work. In Section III, we discuss diacritics in languages using the Latin script. Our approaches to construct chatbots which have the ability to handle misspelled words and words with no diacritics are described in Section IV. The results of our experiments and discussion are presented in Section V. Section VI concludes the paper.

II. LITERATURE REVIEW

This section summarizes common neural network approaches to construct generative chatbots, and then discusses frameworks used to build virtual assistants in the university environment. We also present the benefits and drawbacks of ChatGPT in education, which motivate our selection of Transformer-based models for building a virtual assistant instead of using ChatGPT. We discuss the suitability of Google Bard and Bing Chat for building student-oriented chatbots as well.

A. GENERATION-BASED NEURAL NETWORK CHATBOTS

Common approaches to construct chatbots use deep learning in the form of sequence-to-sequence (seq2seq) models such as LSTMs [3], [4], [5], and Bidirectional RNNs [6]. Additional seq2seq models include the evolved Transformer [7], seq2seq with attention [8], reinforcement learning for seq2seq [9], knowledge graph and hierarchical bi-directional attention [10], and deep seq2seq with GRU cells [11]. Since Transformer [12] is the state-of-the-art model in natural language processing, several chatbots have been constructed using Transformer-based models such as Generative Pre-trained Transformer (GPT) [13], [14], Transformer models trained on augmented human data [15], Deep Bidirectional Transformer or BERT [16], BERT with Knowledge Graph [17], and BERT with Google Dialogflow [18]. BlenderBot,[1] based on a combination of retrieval and generation approaches [19] using the knowledge inside and outside the content of the conversation, outperforms existing chatbots such as Meena [7] and DialoGPT [14].

B. FRAMEWORK-BASED CHATBOTS

Chatbots can be constructed by using chatbot platforms. Social assistants named LiSA [20] and APU Admin Bot [21] were constructed using the Chatfuel[2] chatbot platform to help students in university living, and administrative and academic issues, respectively. The commercial Chatfuel platform uses a rule-based approach for pattern recognition for mapping input questions to outputs through an artificial intelligence model. Chatfuel is inflexible in conversation flows and does not support the use of any local knowledge [22]. Although Chatfuel is the most commonly used platform, it allows users to build straightforward chatbots[3] only.

Barus and Surijati [23] built an assistant for Frequently Asked Questions services in Matana University Library in Indonesia using Dialogflow,[4] focusing on core NLP. The chatbot does not provide correct responses when users mix languages, or use abbreviations or synonyms. An English-Arabic chatbot for university admission [24] was built using the Dialogflow Essentials[5] platform, based on BERT-based models [25]. Dialogflow is easy for users to create chatbots and supports NLP tasks very well [26].

The Microsoft Bot Frameworks[6] provide an all-purpose open-source chatbot platform with two main components, including Bot Builder SDK for developing bots, and Bot Connector Service for relaying information between bots and channels. The Microsoft Bot Frameworks have been used to construct a career counseling chatbot to support Cluj-Napoca's students at Technical University in Romania [27], and to build an undergraduate admission chatbot for the Pamantasan ng Lungsod ng Maynila [28].

Another common open-source framework, called Rasa [29], has been used to construct chatbots for admission [30], university inquiries [31], [32], and answering questions regarding school, academic regulations, and classes [33]. The Rasa framework consists of two main components: Rasa NLU, that uses neural networks, for classifying intents,

[1] https://parl.ai/projects/blenderbot2/
[2] https://chatfuel.com/
[3] https://www.rootinfosol.com/chatbot-development-chatfuel-platform
[4] https://cloud.google.com/dialogflow/docs
[5] https://cloud.google.com/dialogflow
[6] https://dev.botframework.com/

extracting entities, and understanding user messages; and Rasa Core for handling the contexts of conversations using a probabilistic model. Singh and Singh [34] performed experiments on building chatbots for the Central University of Punjab, Bathinda, to analyze the effectiveness of the Rasa and Dialogflow chatbot frameworks. The authors claim that 63% users found that the chatbot built using Rasa was more effective than the one created using Dialogflow.

C. ChatGPT

Since late 2022, it may be claimed that ChatGPT,[7] a text-based chatbot based on a generative pre-trained Transformer GPT model, has dominated discussion of chatbots. ChatGPT has rapidly gained attention for its potential applications in a variety of fields such as healthcare [35], tourism [36], marketing [37], agriculture [38], cosmetic surgery [39], communications [40], and education [41].

The development of education-oriented chatbots also has benefited significantly from ChatGPT [42], [43], [44]. ChatGPT-3 has been impressive, receiving a B on the final exam of an MBA course [45], obtaining a C+ on law school exams [46], and performing at the level of third-year medical students [47]. However, Wood et al. [48] reported that chatGPT ''performed better in answering true/false and multiple-choice questions'' and ''struggled with workout and short-answer questions, with accuracy rates of 28.7 percent and 39.1 percent, respectively''. The majority of questions about programs, staff, academic regulations, and examination schedules are neither true/false nor multiple-choice questions, the strengths of ChatGPT.

ChatGPT-3 was trained on 175 billion parameters, whereas the latest version, ChatGPT-4, is trained on 100 trillion parameters.[8] ChatGPT was trained on a huge general-purpose corpus and uses reinforcement learning from human feedback [46]. ChatGPT has potential limitations, including inaccurate or biased responses because of training on low quality data [49], security issues due to the storage of personal information and sensitive data on the ChatGPT cloud [50], and the cost of required hardware and software, maintenance, and support [49]. ChatGPT needs to be supplied with specific datasets, which may include sensitive information such as personal information or grades, to allow ChatGPT to respond accurately in a university context. This causes serious security and privacy concerns for academic institutions due to data storage in the GPT cloud. Based on our best knowledge, ChatGPT has not released any documents regarding how it collects and stores data [51], and as a result, it is unlikely to receive any kind of data security certifications.[9]

Kocón et al. [52] have discovered that ChatGPT is less stable and performs worse than other SOTA models in almost all tasks. Last but not least, ChatGPT is not accessible from Vietnam. As a result, ChatGPT is not a good choice for our particular situation. We have decided to build a generative virtual assistant to support our students in a specific university in Vietnam.

D. GOOGLE BARD AND BING CHAT

In February 2023, Google AI launched a chatbot, named Bard,[10] based on Google's Language Model for Dialogue Application. Bard is available in 3 languages, US English, Japanese, and Korean, in over 180 countries. Bard [53] was not only trained on a massive dataset but is also able to search the web in real-time.[11] Microsoft has released Bing Chat, which is based on the GPT-4 and integrated into a search engine.[12] Rudolph et al. [54] performed experiments to evaluate the performance of ChatGPT, Google Bard, and Bing Chat. They concluded that none of these AI chatbots reach the level of A-students or B-students. ChatGPT-3.5 performed better than ChatGPT-4 on some questions. Bing Chat and Google Bard tended to get an F-grade on average.

III. DIACRITICS IN LANGUAGES USING LATIN SCRIPT

Our objective is to develop an architecture for training chatbots in languages that use the Latin script with ample use of diacritics. A list of such languages and characters with diacritics (e.g., Vietnamese, Romanian, and French) can be found in Wikipedia.[13] To the best of our knowledge, users may type diacritics or accents of a specific language using a keyboard specifically designed for that language, or an online keyboard[14] or a QWERTY keyboard adapted to that language. Here are some examples:

- Vietnamese[15]: ''aw'' for ''ă'', ''oo'' for ''ô'', or ''as'' for ''á'',
- Romanian[16]: opening bracket ''['' for ''ă'', closing bracket '']'' for ''î'', or semicolon '';'' for ''ş'',
- French[17]: ''′e'' for ''é'', ''″o'' for ''ö'', or ''^e'' for ''ê''.

One of our goals is to investigate Transformer-based models for diacritic restoration and spelling correction for languages using the Latin script with diacritics. We notice that a character with an accent mark or a diacritic is usually created by a combination of a main character and another character or a symbol. Among 23 languages using diacritics Náplava et al. [55] studied, Vietnamese has the most words with diacritics (88.4%) and the highest word error rate compared to a dictionary baseline (40.53%), whereas other languages have fewer than 52.5% words with diacritics and less than 9% word error rate, except for Romanian with 29.71% word error rate. In addition, due to Vietnamese being an Asian language of a lower-middle income country, the

[7] https://chat.openai.com
[8] https://sensoriumxr.com/articles/what-is-chatgpt4
[9] https://www.comm100.com/blog/higher-ed-beware-10-dangers-chatgpt/
[10] https://bard.google.com/
[11] https://zapier.com/blog/chatgpt-vs-bard/
[12] https://zapier.com/blog/chatgpt-vs-bing-chat/
[13] https://en.wikipedia.org/wiki/Diacritic
[14] https://www.lexilogos.com/english/index.htm
[15] https://www.vietnamesepod101.com/blog/2020/10/16/how-to-type-in-vietnamese/
[16] https://www.romanianpod101.com/blog/2020/10/16/how-to-type-in-romanian/
[17] https://www.thoughtco.com/how-to-type-french-accents-1372770

amount of resources devoted to it is substantially lower than other diacritizied languages using the Latin script. Therefore, we perform experiments and evaluate our proposed approach with Vietnamese. Three of the authors are native speakers of Vietnamese, and all authors work in academia, making the design, development, and testing of a student-oriented chatbot in an ideal testbed.

In Vietnamese, using a wrong diacritic may change the meaning of a word. For example, the verb "nghĩ" means "think", while the verb "nghỉ" means "take a rest". Several keyboards such as Unikey[18] and Vietkey[19] are used to input Vietnamese words. There is another complication with Vietnamese orthography. A syllable is the smallest meaning part of Vietnamese orthography [56]. Therefore, Vietnamese words are usually written with syllables separated by white spaces. Depending on the task to be performed in Vietnamese, some approaches may split words based on syllables using white spaces, whereas some specialized Vietnamese segmentation toolkits such as VnTokenizer[20] and Underthesea[21] segment words into syllables separated by white spaces, but clearly indicate the word boundaries. For example, the phrase "công nghệ thông tin" ("information technology") can be segmented into 4 syllables without word boundaries {công, nghệ, thông, tin} by using white spaces as separator in the traditional way or 2 words with separated syllables {công nghệ, thông tin} by using a specialized Vietnamese segmentation toolkit. In our work, we use the methods for inputting Vietnamese words to improve the usability of the chatbots constructed. In addition, we experiment with word segmentation methods using white spaces and the Underthesea toolkit (UTS).

IV. PROPOSED APPROACH

This section presents the generative method used to construct a chatbot in Vietnamese, in both closed and open domains. A closed domain chatbot can answer questions in a specific domain (or area) only, whereas an open domain chatbot can answer questions in any area. We describe the general architecture first. Next, we introduce approaches to handle misspelled words and missing diacritics in questions. Then, we combine all models, and finally link the two chatbots in closed and open domains.

A. OVERVIEW

The overall architecture for the chatbot is given in Figure 1. To handle misspelled words as well as words without diacritics, the architecture uses a Transformer-based submodel trained on especially created datasets, as described in Section IV-B. The goal is to make the chatbot resilient to inputs that are misspelled as well as inputs that do not use diacritics in a highly diacriticized language. The architecture also shortens the questions by removing unimportant words;

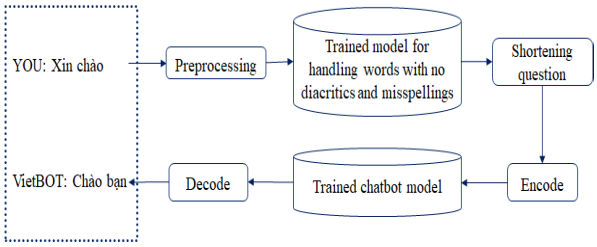


FIGURE 1. The architecture of the chatbot.

this module is placed in the processing pipeline after the text has been standardized by correcting misspellings and restoring diacritics. The main part of the architecture is comprised of the encoder-decoder set-up trained on chat datasets. The datasets and the training of the chat module are discussed later.

B. MODELS FOR HANDLING QUESTIONS WITH NO DIACRITICS AND MISSPELLINGS

Náplava et al. [55] noted that users sometimes type words without diacritics because of the cumbersome requirements of the language inputting software, frequent code switching between English and another language, or due to the need for expedited input compared to typing words with diacritics. In spite of the greatest percentage of words with diacritics and the highest word error rate compared to dictionary baseline as discussed in the previous section, we notice that to search for information faster, Vietnamese people in general and Vietnamese students in particular, prolifically type words with no diacritics. For example, users type "ma hoc phan vi tich phan a1" instead of typing "mã học phần vi tích phân a1" (class code of the calculus a1 course). Moreover, users often forget to turn on the necessary keyboard when typing words with diacritics. For example, in Vietnamese, a user may want to type the word "chào" ("hi"), but it turns into the word "chafo", "chaof", "cha2o", or "chao2". Therefore, we build a Transformer-based model to convert questions having words with no diacritics to questions comprising words with correct diacritics, as presented on the left-hand side of Figure 2.

We create a confusion dataset, named QSet-NoDiacritics, similar to the question set (the QuestionSet dataset), then remove diacritics written above or below the vowels of words. We note that QuestionSet includes questions consisting of words with diacritics and correct spellings. The QuestionSet and QSet-NoDiacritics datasets are fed to the Transformer model. The process of training the Transformer model to convert questions consisting of words without diacritics to questions comprising words with diacritics is similar to the process of training chatbots.

Users may make typographical errors when entering questions, and this affects the quality of the chatbot. Studies have proposed approaches for diacritic restoration using bidirectional RNN for 23 languages using Latin script with diacritics [55], Gated-Recurrent Units for Arabic [57], BERT for Czech [58], and Bidirectional GRU for Vietnamese [59].

[18] https://www.unikey.org/en/
[19] http://vietkey.com.vn/
[20] https://vlsp.hpda.vn/demo/?page=resources
[21] https://github.com/undertheseanlp/underthesea

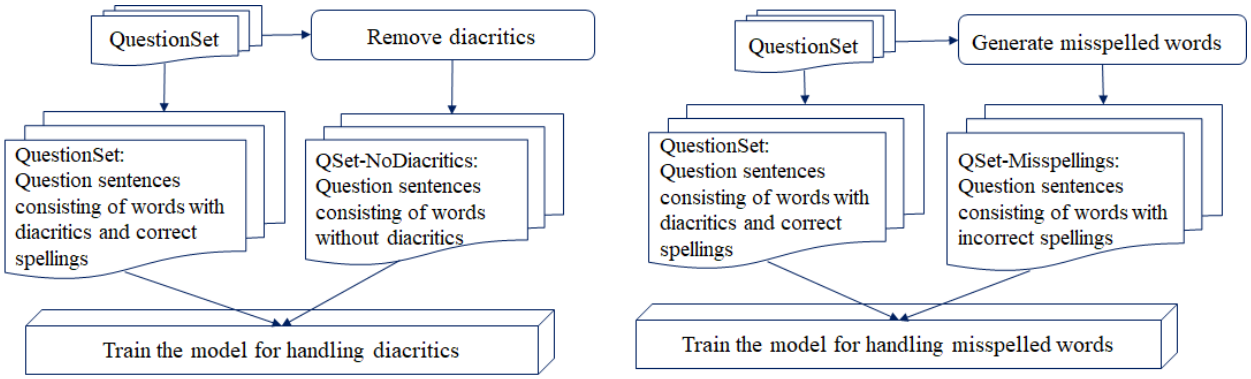


FIGURE 2. The Transformer-based sub-models for improving the chatbot. The left-hand side is the model for handling questions with no diacritics. The right-hand side is the model for transforming questions having words with incorrect spellings to questions having words with correct spellings.

We propose two approaches using the kNN algorithm, as presented by Lam et al. [33], and a Transformer-based model, as shown on the the right-hand side of Figure 2. We construct another confusion dataset, named QSet-Misspellings, consisting of questions with misspellings. In particular, for every sentence in QuestionSet, we randomly create 15 new sentences with misspelled words such that each new sentence contains a character error rate of 20%. The misspelled words in the 15 new sentences are completely different.

At first, we separate the two models, a model for handling diacritics and a model for handling misspellings, for convenience in evaluating their performance and comparing them with other existing models. Then, these two models are integrated to handle questions with no diacritics and misspellings. In particular, the two confusion datasets, QSet-NoDiacritics and QSet-Misspellings, are combined and used to train the Transformer model to transform questions to the correct form with diacritics and correct spellings.

C. SHORTENING QUESTIONS

The questions, obtained after restoring diacritics and correcting misspellings, can be immediately fed to the Transformer-based chatbot to generate responses. However, to help the chatbot understand questions better, as shown in Table 3, we remove unimportant words from the input questions. This task is only applied to the closed domain chatbot. To correctly remove unimportant words from the real data, words in the question set are tokenized and manually labeled. For the university domain, we create 16 labels, some of which are presented in Table 1.

After the steps of diacritic restoration and misspelling correction, questions are tokenized and unimportant words labeled O are removed. For example, a given question "bot ơi, bot à mã học phần ct178 là gì" (hello bot, dear bot, what is the class code ct178) is shortened to become "mã học phần ct178" because the words "bot ơi", "bot à", "là", and "gì" are in the list of words assigned a label of "O". For future work, we will take the advantage of these words labeled to automatically enrich the training dataset and generate diverse answers by discovering semantic relations

TABLE 1. Some labels constructed and their descriptions.

Label	Example values	Description
acronym	VTPA1, Ths, TS,...	Acronyms
position	Dean, vice dean, ...	Position of staff
cohort	cohort, 42, K41	Cohort
group	1, 2, 3, m01, m02,...	Groups of classes
mark	A, B+, B, ...	Final grade
mhp	CT, TN, XH, ...	Class codes
regulation	graduation, tuition fee,...	Academic regulation
verb	pay, withdraw,...	Action
O	stop words and unimportant words	System does not need to pay attention to this word

using a WordNet such as synsets and antonyms. For example, in the Princeton Wordnet, the offset 00744916-a consists of 2 members "difficult" and "hard" having the same meaning of "not easy; requiring great physical or mental effort to accomplish or comprehend or endure". This synset in our Vietnamese WordNet [60], which has the same structure as the Princeton WordNet, also has 2 members "khó" and "gay go" with a gloss "không dễ;...". Therefore, instead of providing an answer "CT178 khó nha bạn" (CT178 is difficult), the chatbot may answer "CT178 gay go nha bạn" (CT178 is hard) or "CT178 không dễ nha bạn" (CT178 is not easy), which also has a similar meaning.

D. COMBINING MODELS AND LINKING CHATBOTS

Our core chatbot model is based simply on the original Transformer [12] following the encoder-decoder framework. After constructing separate models for handling different tasks, we combine all models, as shown in Figure 1. An input question is passed through the pre-processing step. Next, we feed the sentence to the trained model to transform the question to correct form, including diacritic and spelling correction, followed by shortening of questions. The shortened question is considered normal input to the trained chatbot model to generate a corresponding answer.

The architecture for the chatbot in Figure 1 is used to train chatbots individually on closed and open domains using relevant datasets. Input questions are classified using the kNN algorithm, and the correct chatbot domain is chosen to answer the queries. The two chatbots complement each other such

that the closed domain chatbot provides factual information and the open domain chatbot serves as a virtual friend.

V. EXPERIMENTS

A. CONSTRUCTING DATASETS

The architecture presented in Figure 1 is general and will work with any language, in particular a language that uses the Latin script with diacritics. To train VietBOT, we need specific Vietnamese datasets for specific universities. Each dataset used to train the chatbot models consists of one text file for questions and another text file for answers with the constraint that the n^{th} sentence in the first file corresponds to the n^{th} sentence in the second file.

Our specific VietBOT will answer questions primarily inside and secondarily outside the domain of the College of Information Technology and Communication[22] (CICT) of Can Tho University[23] (CTU) in Vietnam. Playing an important role in multi-disciplinary education and training in the Mekong Delta region of Vietnam, CTU has 46,881 students, and CICT in particular has more than 4,045 students in 2022. A chatbot like VietBOT is absolutely necessary to assist a large number of students with information regarding the College, instructors, academic regulation, programs, classes, and so on.

We construct two training datasets, including the CTUBot dataset used for a closed domain chatbot and the OpenSubtitle dataset in Vietnamese, called OpenSubViet, used for an open domain chatbot. The CTUBot dataset is built to help our virtual assistant respond to questions related to CICT, whereas the OpenSubViet dataset is built to help the chatbot communicate with students flexibly and to answer questions in an open domain.

- The CTUBot dataset includes 35,702 question-answer pairs collected from different resources such as the CICT website,[24] academic regulations for students at CTU, bachelor programs of CICT, handbooks for CICT students, and examination schedule for each semester.
- The OpenSubViet dataset comprises over 419,712 dialogue pairs of characters in movies extracted from the OpenSubtitle[25] dataset and translated to Vietnamese using a pre-trained Transformer model.[26] The Transformer translation model was trained on the dataset using 600,000 sentences extracted from TED.[27]

In addition, CICT students were requested to construct a test dataset manually for evaluating VietBOT. This test dataset, named CTUBot-TestSet, comprises 100 questions. Each question has one answer written in 3 different ways. An example of a question and its answers is as follows.

Question: Không học học phần điều kiện được không? (Is it possible to not take the compulsory classes?)
Answer 1: Không nhe bạn (No, it is not possible to not take the compulsory classes.)
Answer 2: Bắt buộc học nhe bạn (You must take the compulsory classes)
Answer 3: Phải học nhe bạn (You have to take the compulsory classes)

TABLE 2. Statistics on the number of words in sentences and dictionaries constructed.

Segment words	Number of words	CTUBot		OpenSubViet	
		Q	A	Q	A
white spaces	Longest sentence	25	42	44	53
	Shortest sentence	4	4	4	3
	Average length	12.54	8.1	9.64	9.59
	Dictionary size	2,742		13,126	
UTS toolkit	Longest sentence	22	39	44	53
	Shortest sentence	3	3	4	3
	Average length	11.68	8.1	9.64	9.6
	Dictionary size	3,349		50,769	

B. DATASET PRE-PROCESSING

Several steps are used to pre-process the training datasets. We convert sentences to lowercase and remove special characters (e.g., @, #, $,...), punctuation marks, and sentence separators (e.g., -, !, ?, ...). We eliminate tokens that have no meaning (e.g., "Hmm" and "ak"). The lines that specify the subtitle author, such as "phim được phụ đề bởi BinhKan" ("movies subtitled by BinhKan"), are removed. Next, acronyms and abbreviations in the answer files are converted to original phrases, such as converting "cntt" to "công nghệ thông tin" ("information technology") or "ths" to "thạc sĩ" ("Master").

A starting markup "<s>" and an ending markup "</s>" are added to the beginning and to the end of every sentence. We segment words in datasets using 2 methods of using white space as a separator and the UTS Vietnamese word segmentation toolkit, and construct a dictionary. Table 2 shows statistics on the number of words in sentences in the datasets. The maximum length of question sentences in the CTUBot dataset is 25. Most questions in the OpenSubViet dataset have a maximum number of 25 words.

C. EXPERIMENTAL RESULTS

We build the models in the Google Colab environment with 12GB RAM with TPU. The hyper-parameters during the training process are as follows: the number of identical layers in encoder and decoder: 6, the number of sub-layers in each layer: 2, the number of epochs: 200, training batch size: 512, dimensionality of input and output: 256, the number of heads: 8, units: 512, a dropout rate: 0.1, an Adam optimizer.

The BLEU metric [61] is used to evaluate the models. The weights of different N-grams used to calculate the BLEU scores of our chatbots are the same as those used by Jason Brownlee[28]: weights of BLEU-1 = (1; 0; 0; 0); weights of

[22] http://www.cit.ctu.edu.vn/encict/
[23] https://en.ctu.edu.vn/
[24] http://www.cit.ctu.edu.vn/
[25] https://www.opensubtitles.org/en/search/subs
[26] https://github.com/pbcquoc/transformer
[27] https://www.ted.com/
[28] https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

TABLE 3. The BLEU scores of VietBOT on the closed domain.

Score	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Not shortening questions	0.525	0.509	0.497	0.451
Split+shortening questions	0.529	0.526	0.517	0.477
UTS+shortening questions	**0.585**	**0.573**	**0.561**	**0.513**

TABLE 4. The BLEU scores of VietBOTs, using the pre-trained embeddings, on the closed domain.

Pre-trained embedding provided by	BLEU-1	BLEU-2	BLEU-3	BLEU-4
PhoBERT	0.520	0.512	0.509	0.462
BARTpho	0.574	0.538	0.503	0.446

BLEU-2 = (0.50; 0.50; 0; 0); weights of BLEU-3 (0.33; 0.33; 0.33; 0) and weights of BLEU-4 = (0.25; 0.25; 0.25; 0.25). In the remainder of this section, we report the experimental results of the Transformer-based chatbot built, the Transformer models to help the chatbot handle questions with no diacritics and misspellings, and finally the performance results of the whole chatbot system.

1) THE CHATBOT MODEL BASED ON TRANSFORMER

The BLEU scores of VietBOT in the closed domain are presented in Table 3. The results show that shortening questions helps the chatbot achieve better results. For example, given a question "bot ơi, bot à mã học phần ct178 là gì" ("hello bot, dear bot what is the class code ct178"), the chatbot without shortening questions could not understand the question well, so that it responds with "CT178". The chatbot with shortened questions recognizes the question entities correctly, and it provides a correct answer "Nguyên lý hệ điều hành 3 tín chỉ" (Principles of Operating System 3 credits). The chatbot that segments words using the UTS toolkit gets higher BLEU scores than the one using white space as a separator, named Split.

We also experiment with the pre-trained BPE embeddings provided by PhoBERT [62], which is trained on 20GB of Vietnamese text. We simply map words in our dictionary with words provided by PhoBERT, and extract the corresponding embeddings of words. Because the size of pre-trained BPE embeddings is 768, the SVD factorization is used to reduce their size to 256, which is accepted by our Transformer model. Similarly, we experiment with the pre-trained embeddings supported by BARTpho [63]. Table 4 presents the BLEU scores of VietBOT using pre-trained embeddings supported by PhoBERT and BARTpho. VietBOT using our trained embeddings achieves the highest BLEU scores among VietBOTs using pre-trained embeddings.

We do not shorten questions on the open domain chatbot because we do not know the exact possible entities in open domain questions. The BLEU-1, 2, 3, and 4 scores of the chatbot on the open domain are 0.18, 0.13, 0.10, and 0.08, respectively. The BLEU-4 score of our open domain chatbot is 0.08, which is slightly better than the chatbot constructed using the reinforcement learning approach and the LSTM model (which is 0.07) on the same dataset in Vietnamese [64]. Our model needs about 3 hours for training, while Nguyen's model [64] needs about 72 hours on the same kind of hardware. This shows that the learning speed of the Transformer model is better than the reinforcement learning and the LSTM.

TABLE 5. The BLEU scores of the Transformer-based models used to handle questions with no diacritics, misspellings, and both no diacritics and misspellings (the column "Neither").

Domain	Scores	No diacritics	Misspellings	Neither
Closed	BLEU-1	0.89	0.84	0.82
	BLEU-2	0.85	0.79	0.76
	BLEU-3	0.82	0.75	0.72
	BLEU-4	0.76	0.69	0.66
Open	BLEU-1	0.98	0.98	0.98
	BLEU-2	0.98	0.98	0.97
	BLEU-3	0.97	0.97	0.95
	BLEU-4	0.95	0.94	0.92

2) MODELS FOR HANDLING QUESTIONS WITH NO DIACRITICS AND MISSPELLINGS

Both the CTUBot and OpenSubViet datasets are used to train the model to convert questions with typos and no diacritics to questions with correct spellings. The BLEU scores of the tasks for handling questions having words with no diacritics, mistyped words, and both these issues at the same time using the Transformer-based model are presented in Table 5.

Table 6 shows some examples of the Transformer models used to correct mistyped words and words with no diacritics. The first two mistyped questions are converted correctly. In the last question, the model also performs correction well, except that there is one incorrectly transformed word "biết". In this question, the user actually wants to know the process for requesting an I grade (the word "xin"), but the model transforms it to simply want to know information about an I grade (a status of "Incomplete" the class) in general.

To compare, we calculate the accuracy of our Transformer-based models for handling misspelled words and missing diacritics. Our Transformer model gets accuracies of 82% and 76% for the tasks of handling sentences without diacritics, and sentences with mistyped words, respectively. We also perform experiments to improve the chatbot model by using the kNN algorithm and Euclidean distance as suggested by Lam et al. [33]. The values of k are odd numbers from 1 to 17. The kNN algorithm for handling questions having misspelled words achieves the highest accuracy of 72% with $k = 1$. We notice that the Transformer model corrects misspelled words better than the kNN. Therefore, we use the Transformer model to handle this task in the VietBOT.

3) COMBINING MODELS

Finally, we integrate the models to handle questions with no diacritics and misspellings into the chatbot model. Table 7 shows a conversation between a user (YOU) and VietBOT. The "Meaning" column shows the actual question a user wants to ask when the user makes typos, and the content

TABLE 6. Some examples of handling words with no diacritics and mistyped words.

User wants to ask	User typed	Result of the model for correction
CT205 mở lớp ở học kỳ nào? (Which semester is the CT205 class offered?)	gt205 mở lớp o học kỳ nào	ct205 mở lớp ở học kỳ nào
Học phần tiên quyết của toán rời rạc là gì? (What is the prerequisite class of the Discrete Mathematics class?)	hok phần tiên quyết của toán rời rạc là j	học phần tiên quyết của toán rời rạc là gì?
Nếu tôi muốn xin điểm I thì phải làm sao? (If I want to apply for an I grade, what should I do) Note: An I grade means the student has not completed the class.	nếu tôy muốn xin điểm I thi2 fai làm sao	nếu tôi muốn biết điểm I thì phải làm sao (If I want to know about an I grade, what should I do)

TABLE 7. Example of conversation between a user (YOU) and the VietBOT.

Who	Q/A	Meaning
YOU	bot biết co Khang khong	bot biết cô Khang không - (Do you know Ms. Khang)
VietBOT	Mình biết nhe, là cô Lâm Nhựt Khang ở bộ môn CNTT	(I know, she is Ms. Lam Nhut Khang in the Department of IT)
YOU	truong khoa công ngheej thông tin là ai	trưởng khoa công nghệ thông tin là ai (Who is the Dean of the College of Information Technology)
VietBOT	Trưởng khoa công nghệ thông tin là thầy Nguyễn Hữu Hòa	(The dean is Mr. Nguyen Huu Hoa)
YOU	ma hoc phan ltcb a	mã học phần ltcb a (class code of ltcb a. "ltcb a" is an acronym of "lập trình căn bản A" which means "Basic Programming A")
VietBOT	CT101	CT101. Note: VietBot understands the acronym "ltcb" and reposes correctly.
YOU	lịch thi học kỳ 2 2020 2021 ct112 nhóm 8	(examination schedule of ct112, group 8, semester 2, school year 2020-2021)
VietBOT	lịch thi là 25/5/2021, phòng thi là P12	(the examinations are on 25/5/2021, in room P12)

inside the parentheses ( ) is the meaning of the actual questions in English.

In addition, we generate noisy questions by randomly removing diacritics or replacing correct words with misspelled words from the CTUBot-TestSet and evaluate the robustness of VietBOT through its responses, as shown in Table 8. The percentages in the columns of "No diacritics" and "Misspellings" in Table 8 present the error rates in user questions.

D. DISCUSSION

Our Transformer-based model to support the chatbot system consists of a word diacritic restoration model and a mistyped word correction model. It is not easy to compare the performance of different approaches on diverse datasets using various evaluation metrics. For the sake of completeness, we make an attempt at comparing our models with published papers.

First, we compare our proposed model with existing models used to handle Vietnamese words with no diacritics. Tran et al. [59] use several models to solve the diacritic restoration problem. The experimental results show that the Bidirectional-GRU-RNN is the best model with a BLEU score of 88.06%, whereas the BLEU scores of the GRU-RNN and Bidirectional-LSTM-RNN models are 73.47% and 87.98%, respectively, on their dataset. Our Transformer-based model for word diacritic restoration in both closed and open domains is better than the best model of Tran et al. [59].

Second, we compare our Transformer-based model used to correct Vietnamese mistyped words with other approaches. The N-gram model [65] based on the contexts on both sides of syllables is currently the best approach for spelling correction with 94% precision on their dataset. The LSTM model used for Vietnamese consonant misspelling correction achieves 90.56% accuracy on sentences [66], whereas the N-gram model [65] reaches 72.18% accuracy. Do et al. [67] use the Transformer model to correct mistyped and misspelled errors. Each syllable is used to generate a candidate list of common types of Vietnamese spelling errors. The training datasets are generated by adding errors to non-error sentences extracted from a Vietnamese news corpus.[29] Their model achieves performance with 81.5% errors corrected, while the metric of the N-gram model [65] is 79.3%. Our model is not good as their model. However, Do et al. [67] standardize marks and syllables based on Telex typing form which may not let the model work well with words typed using Vni formats. Our model can handle mistyped and misspelled words created by both Telex and Vni text input formats. Besides, we can save time and effort required to construct the training confusion datasets containing mistyped and misspelled words because our confusion datasets are constructed automatically. A published paper in 2022 on simultaneous correcting diacritics and typos [68] suggests the combination of ByT5 Transformer and the classical dictionary methods can improve the accuracy to 87.86% in Vietnamese. For future work, we will improve our Transformer-based proposed approach to improve the chatbot system by discovering information from Vietnamese dictionaries [68] and taking into account the confusion set for each syllable based on edit distance and chosen Vietnamese characteristics [65].

Our VietBOT performs better than an Indonesian chatbot for university admissions constructed using the seq2seq model without and with attention [8] which achieves BLEU scores of 43.61 and 44.68, respectively. Based on our best knowledge, we have not found many published papers on constructing Vietnamese chatbots. Nguyen and

[29] https://github.com/binhvq/news-corpus

TABLE 8. The BLEU scores of the integrated VietBOT models for handling questions with misspellings and no diacritics.

Option	No diacritics	Misspellings	BLEU-1	BLEU-2	BLEU-3	BLEU-4
No shortening questions	100%	-	0.534	0.520	0.509	0.463
	50%	-	**0.552**	**0.538**	**0.520**	**0.473**
	10%	-	0.535	0.521	0.507	0.459
	-	50%	0.492	0.482	0.471	0429
	-	10%	0.517	0.502	0.496	0.449
	100%	50%	0.478	0.463	0.446	0.402
	50%	50%	0.498	0.475	0.461	0.411
	30%	30%	0.490	0.475	0.465	0.423
	10%	10%	0.532	0.517	0.506	0.462
Split & shortening questions	100%	-	0.578	0.562	0.543	0.494
	50%	-	**0.593**	**0.574**	**0.554**	**0.504**
	10%	-	0.571	0.555	0.537	0.486
	-	50%	0.519	0.508	0.505	0.449
	-	10%	0.539	0.523	0.512	0.464
	100%	50%	0.460	0.447	0.428	0.387
	50%	50%	0.511	0.489	0.471	0.418
	30%	30%	0.511	0.495	0.478	0.435
	10%	10%	0.559	0.544	0.531	0.487
UTS & shortening questions	100%	-	0.561	0.554	0.546	0.503
	50%	-	0.579	0.567	0.554	0.510
	10%	-	**0.581**	**0.574**	**0.563**	**0.516**
	-	50%	0.529	0.253	0.517	0.477
	-	10%	0.563	0.552	0.547	0.500
	100%	50%	0.498	0.488	0.477	0.437
	50%	50%	0.503	0.488	0.477	0.436
	30%	30%	0.508	0.494	0.485	0.446
	10%	10%	0.562	0.553	0.546	0.505

Shcherbakov [69] constructed a Vietnamese chatbot using a seq2seq model with an attention decoder. They claim that the BLEU score of their "whole model is 1.443%". Some Vietnamese chatbots have been created using the Rasa framework. The NEU-chatbot [30], created using Rasa, for admission to National Economics University in Vietnam answers 90.29% questions and achieves an accuracy of 97.1% on their test set. None of the chatbots discussed [8], [30], [69] have a component to handle misspellings or lack of diacritics. The chatbot of Lam et al. [33], called CTU-chatbot, also created using Rasa, for students at Can Tho University in Vietnam responds 92.78% questions and reaches the best accuracy of 94.33%. It is unfair for us to compare straightforwardly our chatbot with these chatbots created using the Rasa framework. Therefore, we perform experiments with our VietBOT and CTU-chatbot on our dataset. The results show that VietBOT responds to questions of actual users with an accuracy of 73%; while CTU-chatbot only answers questions with an accuracy of 37%. The BLEU scores of CTU-chatbot are BLEU-1: 0.304, BLEU-2: 0.311, BLEU-3: 0.333, and BLEU4: 0.309, which are lower than VietBOT in all cases. CTU-chatbot cannot answer questions which are not in the training dataset, whereas VietBOT can. Moreover, constructing a training dataset for the chatbot of Lam et al. [33] costs more time and effort than for our VietBOT. While performing experiments, we intend to collect questions from users to enrich question files for training VietBOT; then, we may extract answers automatically from documents or create answers manually. For future work, VietBOT is more suitable for expansion than the chatbot of Lam et al. [33].

Normally, universities do not have a bank of question-answer pairs for training a chatbot. The training datasets may be obtained from different resources as follows.

- Individual staff of the university, such as instructors, advisors, and secretaries, may be able to provide the highest quality question-answer pairs for training the chatbot because they are the most familiar with the school, faculties, classes, school regulations, and so on. However, this method is likely to be prohibitively expensive and time consuming, if at all possible due to lack of interest or lack of time. In our work, advisors and instructors helped contribute some question-answer pairs manually.
- University websites and published documents of a university, such as handbooks and academic regulations, provide reliable data. For future work, we will investigate approaches to automatically generate question-answer pairs from these sources using approaches such as the OneStop model [70] based on the canonical seq2seq Transformer.
- We may extract question-answer pairs from the support applications provided by a university [8] using Whatsapp and Zalo, and Fanpage on Facebook of the university [30], [71].

VI. CONCLUSION

We have developed a virtual assistant for students based on the generation-based model using the attention-only Transformer neural network model. Our method can construct a chatbot for a university which has a collection of question-answer pairs. Our Transformer-based models are simple but effective for the chatbots to handle questions that lack diacritics and have misspellings in Vietnamese. Although we have not performed experiments using our proposed models on datasets in other languages, we believe our models could contribute to improve chatbots in other languages which also

use Latin script and have diacritics. In actual experiments with real users, VietBOT can answer questions not in the training dataset, understand questions in which users have misspelled or typed without diacritics, and respond to users with relevant answers. We are improving our virtual assistant system by exploring approaches to automatically enlarge the training dataset, providing more functions to meet the needs of students.

REFERENCES

[1] G. Caldarini, S. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots," *Information*, vol. 13, no. 1, p. 41, Jan. 2022.

[2] H. Raval, "Limitations of existing chatbot with analytical survey to enhance the functionality using emerging technology," *Int. J. Res. Anal. Rev. (IJRAR)*, vol. 7, no. 2, pp. 7–10, 2020.

[3] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 110–119.

[4] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 994–1003.

[5] Z. Yin, K.-H. Chang, and R. Zhang, "DeepProbe: Information directed sequence understanding and chatbot design via recurrent neural networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 2131–2139.

[6] M. Dhyani and R. Kumar, "An intelligent chatbot using deep learning with bidirectional RNN and attention model," *Mater. Today, Proc.*, vol. 34, pp. 817–824, Jan. 2021.

[7] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," 2020, *arXiv:2001.09977*.

[8] Y. W. Chandra and S. Suyanto, "Indonesian chatbot of university admission using a question answering system based on sequence-to-sequence model," *Proc. Comput. Sci.*, vol. 157, pp. 367–374, Jan. 2019.

[9] I. V. Serban, C. Sankar, M. Germain, S. Zhang, Z. Lin, S. Subramanian, T. Kim, M. Pieper, S. Chandar, N. R. Ke, S. Rajeshwar, A. de Brebisson, J. M. R. Sotelo, D. Suhubdy, V. Michalski, A. Nguyen, J. Pineau, and Y. Bengio, "A deep reinforcement learning chatbot," 2017, *arXiv:1709.02349*.

[10] Q. Bao, L. Ni, and J. Liu, "HHH: An online medical chatbot system based on knowledge graph and hierarchical bi-directional attention," in *Proc. Australas. Comput. Sci. Week Multiconf.*, 2020, pp. 1–10.

[11] G. Sperlí, "A deep learning based chatbot for cultural heritage," in *Proc. 35th Annu. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 935–937.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[13] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, "CAiRE: An end-to-end empathetic chatbot," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 9, 2020, pp. 13622–13623.

[14] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT: Large-scale generative pre-training for conversational response generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2020, pp. 270–278.

[15] J. J. Bird, A. Ekárt, and D. R. Faria, "Chatbot interaction with artificial intelligence: Human data augmentation with T5 and language transformer ensemble for text classification," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 4, pp. 3129–3144, 2021.

[16] R. Bathija, P. Agarwal, R. Somanna, and G. B. Pallavi, "Guided interactive learning through chatbot using bi-directional encoder representations from transformers (BERT)," in *Proc. 2nd Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Mar. 2020, pp. 82–87.

[17] S. Yoo and O. Jeong, "An intelligent chatbot utilizing BERT model and knowledge graph," *J. Soc. e-Bus. Stud.*, vol. 24, no. 3, pp. 87–98, 2020.

[18] N. Kanodia, K. Ahmed, and Y. Miao, "Question answering model based conversational chatbot using BERT model and Google Dialogflow," in *Proc. 31st Int. Telecommun. Netw. Appl. Conf. (ITNAC)*, Nov. 2021, pp. 19–22.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttle, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[20] M. Dibitonto, K. Leszczynska, F. Tazzi, and C. M. Medaglia, "Chatbot in a campus environment: Design of LiSA, a virtual assistant to help students in their university life," in *Proc. Int. Conf. Human-Comput. Interact.* Cham, Switzerland: Springer, 2018, pp. 103–116.

[21] J. Singh, M. H. Joesph, and K. B. A. Jabbar, "Rule-based chabot for student enquiries," *J. Phys., Conf. Ser.*, vol. 1228, no. 1, May 2019, Art. no. 012060.

[22] A. Vishwakarma and A. Pandey, "A review & comparative analysis on various chatbots design," *Int. J. Comput. Sci. Mobile Comput.*, vol. 10, no. 2, pp. 72–78, Feb. 2021.

[23] S. P. Barus and E. Surijati, "Chatbot with Dialogflow for FAQ Services in Matana university library," *Int. J. Informat. Comput.*, vol. 3, no. 2, pp. 62–70, 2022.

[24] W. El Hefny, Y. Mansy, M. Abdallah, and S. Abdennadher, "Jooka: A bilingual chatbot for university admission," in *Proc. World Conf. Inf. Syst. Technol.*, 2021, pp. 671–681.

[25] W. El Hefny, A. El Bolock, C. Herbert, and S. Abdennadher, "Towards a generic framework for character-based chatbots," in *Proc. Int. Conf. Practical Appl. Agents Multi-Agent Syst.* Cham, Switzerland: Springer, 2020, pp. 95–107.

[26] P. Kostelník, I. Pisařovic, M. Muron, F. Dařena, and D. Procházka, "Chatbots for enterprises: Outlook," *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, vol. 67, no. 6, pp. 1541–1550, Dec. 2019.

[27] M. Conţ, A. Ciupe, B. Orza, I. Cohuţ, and G. Niţu, "Career counseling chatbot using Microsoft Bot Frameworks," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Mar. 2022, pp. 1387–1392.

[28] A. G. Usigan, M. I. Salomeo, G. J. L. J. Zafe, C. J. Centeno, A. A. R. C. Sison, and A. G. Bitancor, "Implementation of an undergraduate admission chatbot using Microsoft Azure's question answering and Bot Framework," in *Proc. 5th Artif. Intell. Cloud Comput. Conf.*, Dec. 2022, pp. 240–245.

[29] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open source language understanding and dialogue management," 2017, *arXiv:1712.05181*.

[30] T. T. Nguyen, A. D. Le, H. T. Hoang, and T. Nguyen, "NEU-chatbot: Chatbot for admission of national economics university," *Comput. Educ., Artif. Intell.*, vol. 2, Jan. 2021, Art. no. 100036.

[31] Y. Windiatmoko, R. Rahmadi, and A. F. Hidayatullah, "Developing Facebook chatbot based on deep learning using RASA framework for university enquiries," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1077, no. 1, Feb. 2021, Art. no. 012060.

[32] S. Meshram, N. Naik, V. Megha, T. More, and S. Kharche, "College enquiry chatbot using Rasa framework," in *Proc. Asian Conf. Innov. Technol. (ASIANCON)*, Aug. 2021, pp. 1–8.

[33] K. N. Lam, N. N. Le, and J. Kalita, "Building a chatbot on a closed domain using RASA," in *Proc. 4th Int. Conf. Natural Lang. Process. Inf. Retr.*, Dec. 2020, pp. 144–148.

[34] S. Singh and S. Singh, "Effective analysis of chatbot frameworks: RASA and Dialogflow," EasyChair, Manchester, U.K., Tech. Rep. 8338, 2022.

[35] F. Muftić, M. Kadunić, A. Mušinbegović, and A. A. Almisreb, "Exploring medical breakthroughs: A systematic review of ChatGPT applications in healthcare," *Southeast Eur. J. Soft Comput.*, vol. 12, no. 1, pp. 13–41, 2023.

[36] I. Carvalho and S. Ivanov, "ChatGPT for tourism: Applications, benefits and risks," *Tourism Rev.*, Apr. 2023, doi: 10.1108/TR-02-2023-0088.

[37] P. Rivas and L. Zhao, "Marketing with ChatGPT: Navigating the ethical terrain of GPT-based chatbot technology," *AI*, vol. 4, no. 2, pp. 375–384, Apr. 2023.

[38] S. Biswas, "Importance of ChatGPT in agriculture: According to Chat-GPT," Univ. Tennessee-Health Sci. Center, Memphis, TN, USA, Tech. Rep., 2023. [Online]. Available: https://ssrn.com/abstract=4405391

[39] R. Gupta, P. Pande, I. Herzog, J. Weisberger, J. Chao, K. Chaiyasate, and E. S. Lee, "Application of ChatGPT in cosmetic plastic surgery: Ally or antagonist?" *Aesthetic Surg. J.*, vol. 43, no. 7, pp. NP587–NP590, Jun. 2023.

[40] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic importance-aware communications using pre-trained language models," 2023, *arXiv:2302.07142*.

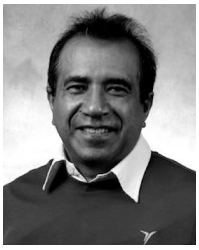
[41] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence: Understanding the potential benefits of ChatGPT in promoting teaching and learning," Queen's Univ., Canada, Tech. Rep., 2023. [Online]. Available: https://ssrn.com/abstract=4337484

[42] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, May 2023, pp. 1–9.
[43] M. Sullivan, A. Kelly, and P. McLaughlan, "ChatGPT in higher education: Considerations for academic integrity and student learning," *J. Appl. Learn. Teach.*, vol. 6, no. 1, pp. 1–11, 2023.
[44] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *J. Appl. Learn. Teach.*, vol. 6, no. 1, pp. 342–360, Jan. 2023.
[45] C. Terwiesch, "Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course," Mack Inst. Innov. Manag. Wharton School, Univ. Pennsylvania, Philadelphia, PA, USA, 2023.
[46] J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz, "ChatGPT goes to law school," *J. Legal Educ.*, Forthcoming. [Online]. Available: SSRN: https://ssrn.com/abstract=4335905
[47] A. Gilson, C. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash, "How does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment," *medRxiv*, pp. 2022–2034, 2022, doi: 10.1101/2022.12.23.22283901.
[48] D. A. Wood et al., "The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions?" *Issues Accounting Educ.*, vol. 38, no. 4, pp. 1–28, Nov. 2023.
[49] J. Su and W. Yang, "Unlocking the power of ChatGPT: A framework for applying generative AI in education," *ECNU Rev. Educ.*, vol. 6, no. 3, pp. 355–366, doi: 10.1177/20965311231168423.
[50] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "ChatGPT: Applications, opportunities, and threats," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, 2023, pp. 274–279.
[51] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learn. Environ.*, vol. 10, no. 1, p. 15, Feb. 2023.
[52] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocon, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Milkowski, M. Oleksy, M. Piasecki, Ł. Radlinski, K. Wojtasik, S. Wozniak, and P. Kazienko, "ChatGPT: Jack of all trades, master of none," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101861.
[53] S. Siad, "The promise and Perils of Google's Bard for scientific research," Int. Centre Adv. Medit. Agronomic Stud., Bari, Italy, Mar. 2023, doi: 10.17613/yb4n-mc79.
[54] J. Rudolph, S. Tan, and S. Tan, "War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education," *J. Appl. Learn. Teach.*, vol. 6, no. 1, pp. 364–389, 2023.
[55] A. Hucko and P. Lacko, "Diacritics restoration using deep neural networks," in *Proc. World Symp. Digit. Intell. Syst. Mach. (DISA)*, Aug. 2018, pp. 195–200.
[56] B. N. Ngo and B. H. Tran, "The Vietnamese language learning framework," *J. Southeast Asian Lang. Teach.*, vol. 10, pp. 1–24, May 2001.
[57] A. Alkhatlan, F. Kateb, and J. Kalita, "Attention-based sequence learning model for Arabic diacritic restoration," in *Proc. 6th Conf. Data Sci. Mach. Learn. Appl. (CDMA)*, Mar. 2020, pp. 7–12.
[58] J. Náplava, M. Straka, and J. Straková, "Diacritics restoration using BERT with analysis on Czech language," *Prague Bull. Math. Linguistics*, vol. 116, no. 1, pp. 27–42, Apr. 2021.
[59] Q.-L. Tran, G.-H. Lam, V.-B. Duong, and T.-H. Do, "A study on diacritic restoration problem in Vietnamese text using deep learning based models," in *Proc. IEEE Int. Conf. Commun., Netw. Satell. (COMNETSAT)*, Jul. 2021, pp. 306–310.
[60] K. N. Lam, T. H. To, T. T. Tran, and J. Kalita, "Improving Vietnamese WordNet using word embedding," in *Proc. 3rd Int. Conf. Natural Lang. Process. Inf. Retr.*, Jun. 2019, pp. 110–114.
[61] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.
[62] D. Q. Nguyen and A.-T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 1037–1042.
[63] N. L. Tran, D. Le, and D. Q. Nguyen, "BARTpho: Pre-trained sequence-to-sequence models for Vietnamese," in *Proc. Interspeech*, Sep. 2022, pp. 1751–1755.
[64] V. Nguyen, "Constructing a chatbot using reinforcement learning approach," Bachelor thesis, Dept. Inf. Technol., Can Tho Univ., Vietnam, 2019.
[65] T. X. H. Nguyen, T.-T. Dang, T.-T, Nguyen, and A.-C. Le, "Using large N-gram for Vietnamese spell checking," in *Knowledge and Systems Engineering*. Cham, Switzerland: Springer, 2015, pp. 617–627.
[66] H. T. Nguyen, T. B. Dang, and L. M. Nguyen, "Deep learning approach for Vietnamese consonant misspell correction," in *Proc. Int. Conf. Pacific Assoc. Comput. Linguistics*. Cham, Switzerland: Springer, 2019, pp. 497–504.
[67] D.-T. Do, H. T. Nguyen, T. N. Bui, and H. D. Vo, "VSEC: Transformer-based model for Vietnamese spelling correction," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2021, pp. 259–272.
[68] L. Stankevičius, M. Lukoševičius, J. Kapočiūtė-Dzikienė, M. Briedienė, and T. Krilavičius, "Correcting diacritics and typos with a ByT5 transformer model," *Appl. Sci.*, vol. 12, no. 5, p. 2636, Mar. 2022.
[69] T. Nguyen and M. Shcherbakov, "A neural network based Vietnamese chatbot," in *Proc. Int. Conf. Syst. Model. Advancement Res. Trends (SMART)*, Nov. 2018, pp. 147–149.
[70] S. Cui, X. Bao, X. Zu, Y. Guo, Z. Zhao, J. Zhang, and H. Chen, "OneStop QAMaker: Extract question-answer pairs from text in a one-stop approach," 2021, *arXiv:2102.12128*.
[71] Y. Windiatmoko, A. F. Hidayatullah, and R. Rahmadi, "Developing FB chatbot based on deep learning using RASA framework for university enquiries," 2020, *arXiv:2009.12341*.

KHANG NHUT LAM received the bachelor's degree in computer science from Can Tho University, Vietnam, in 2005, the master's degree in information technology from Ewha Womans University, South Korea, in 2009, and the Ph.D. degree in computer science from the University of Colorado Colorado Springs, USA, in 2015. Since 2015, she has been a Senior Lecturer with the Department of Information Technology, Can Tho University. Her research interests include natural language processing, question-answering systems, image captioning, and deep learning.

LOC HUU NGUY is currently pursuing the degree with the College of Information and Communication Technology, Can Tho University, Vietnam. He has participated in several research projects in natural language processing and question-answering systems.

VAN LAM LE received the bachelor's degree in informatics from Can Tho University, Vietnam, the master's degree in information technology from The University of Newcastle, Australia, and the Ph.D. degree in computer science from the Victoria University of Wellington. He is a Senior Lecturer with the College of Information and Communication Technology, Can Tho University, where he has been a Lecturer, since 2000. His research is focused on network security, the IoT, digital transformation, and machine learning.



JUGAL KALITA is a Professor of computer science with the University of Colorado Colorado Springs, USA. His research interests include natural language processing, computational linguistics, machine learning, deep learning. He, his students, and collaborators have published over 250 papers with more than 12,000 citations. He has written several books, including the *Machine Learning: Theory and Practice* (CRC Press, 2023).