ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 5, 207-212 (2019)

www.astesj.com

Special Issue on Advancement in Engineering and Computer Science

ASTESJ

ISSN: 2415-6698

Integrating Diacritics Restoration and Question Classification into Vietnamese Question Answering System

Bui Thanh Hung*

Data Analytics & Artificial Intelligence Laboratory, Engineering - Technology Faculty, Thu Dau Mot University, 590000, Vietnam

ARTICLE INFO

Article history:
Received: 30 May, 2019
Accepted: 13 September, 2019
Online: 08 October, 2019

Keywords:
Diacritics Restoration
Question Classification
Vietnamese Question Answering System
Deep Learning
Encoder-Decoder LSTM
Bi-LSTM

ABSTRACT

This paper presents a solution for question answering system for Vietnamese language by integrating diacritics restoration and question classification via deep learning approach. It could be said that this will be the first research integrating two phases into Vietnamese question answering system. Question classification has a critical role in the question answering system. However if the question has too many missing diacritics, this will make the classification extremely more difficult. In this paper, both automatic insertion of diacritics and question classification tasks are built to rely on deep learning approach. For diacritics restoration task, we apply the Encoder-Decoder LSTM model. The result of the first step will be the input of question classification. We use pre-train word embeddings in the Bidirectional LSTM model for Vietnamese question classification. The deep learning approach for both tasks is powerful and highly accurate model. By integrating diacritics restoration and question classification into Vietnamese question answering system – ICTbot of Binh Duong Department of Information and Communications Support System – it has produced remarkably positive results; thus proves the practicability of this proposed system.

1. Introduction

Internet allows the researchers to save data and make them available to the public. However, it also makes the search of information in such big data environment more complicated and expensive. Question Answering systems have been developed as a new advanced research tools for this issue.

An automated question answering system has several key steps to follow such as:

- The user inputs the question through query interface. The question is analyzed, pre-processed and phrased into words.
- Question classification identifies the type of the question. This step is rather critical to achieve the most appropriate answering type.
- Finally, based on the result of Question classification, the system will choose the appropriate answer from the answer sets.

In the first step, if there are too many missing diacritics in questions, it will make it challenging for the second step - Question classification. A diacritic is a mark written either above or below a letter to alter its original phonetic or orthographic value. Diacritics are used in several languages' orthography. Nevertheless, in daily conversation, as a matter of convenience, diacritics are usually ignored in many languages. The absence of diacritics in question text makes it extremely challenging for both Question Classification and Question Answering system.

The Vietnamese alphabet uses Latin script with diacritical signs. However, Vietnamese has seven modified letters including ă, â, đ, ê, ô, ơ, and ư and also six tone marks including unmarked tone (ngang), acute (sắc), grave (huyền), hook above (hỏi), tilde (ngã), and underdot (nặng). Due to its diacritical complexity, when typing on phones or computers, most of Vietnamese people prefer to type non-diacritic characters as a matter of convenience. As a consequence, it sometimes causes the confusion or mis-interpretation for the recipients or readers. With that being said, integrating diacritics restoration into question classification is a critical task for Vietnamese question answering system.

We propose the Vietnamese Question Answering model with two phases in this paper. The first phase is automatically recovering the missing diacritics. And the second phase is integrating diacritics restoration into Vietnamese question

*Bui Thanh Hung, Engineering - Technology Faculty, Thu Dau Mot University, hungbt.cntt@tdmu.edu.vn

answering system. We use the Encoder-Decoder LSTM model for automatically recovering the missing diacritics and the Bidirectional LSTM model for Vietnamese question classification. We integrate diacritics restoration into Question classification of Vietnamese question answering system.

This paper consists of five sections. The related works are reviewed in Section 2. Our method is presented in Section 3. Section 4 mentions the procedure of constructing dataset, the experiment settings and discusses the outcomes of the experiments. Our work and possible future research are concluded in Section 5.

2. Related Words

Firstly, diacritics restoration – as the first step – could be run following two typical approaches: character-based [1-2] and word-based [3-5]. For Vietnamese language, many researches propose their methods to restore accents such as [6-9]. In this research we choose the Encoder-Decoder LSTM model that proposed recently by Bui [6] because of the advanced approach, and high accuracy for Vietnamese diacritics restoration.

For question classification, there are three main approaches: Rule-based, Machine Learning and Hybrid ones [10]. Using a few manually handcrafted rules to match the question is the rule-based approach. One of key challenges here is to define too many rules. Machine-learning approach is to train a classifier and use the trained classifier to predict the class label. Combining the two above - the rule-based and machine learning - is the hybrid approach. In terms of question classification, the most successful approaches are Machine Learning and hybrid methods. Nguyen et al. [11] proposed a SVM based method for the same task. Firstly, the question was parsed and tokenized, parts-of-speech were tagged, data removed stop-words and was stemmed. Next they extracted a lot of features and selected features before passing the data into a support vector machine for training. For test questions they used the same pre-processing. Deep Learning approaches have proved that it is significantly beneficial for text classification, summarization, machine translation, etc. as well as for question classification [12-15]. Andreas et al. [12] proposed compose neural networks for question answering. Kim et al. [16] employed Convolutional Neural Networks (CNNs) and treated questions as general sentences to achieve remarkably strong performance in the TREC question classification task.

With regards to our research, what differentiates it from the others is the Vietnamese question answering system. This is the first research integrating two phases - diacritics restoration and question classification into Vietnamese question answering system. The diacritics restoration is the Encoder-Decoder LSTM model and Question classification uses the Bi-LSTM model. The result of the first step will be the input of question classification.

3. Methodology

Our model includes two steps: Diacritics Restoration and Question Classification. In the first step, we use Encoder-Decoder LSTM model proposed by Bui [6] and in the second step, pre-train word embeddings are used in the Bidirectional LSTM model for Vietnamese question classification. We integrate the result of the first step into question classification. An overview of our architecture is illustrated in Figure 1. We will describe more details of each layer in our model as follows:

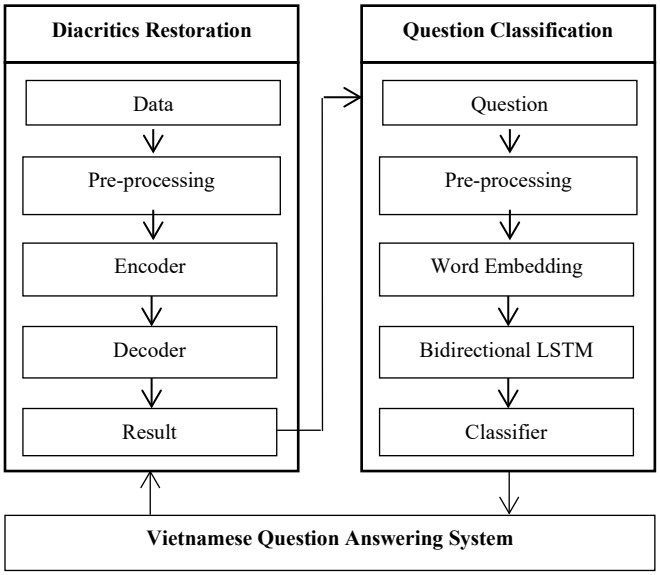

Figure 1: The proposed model

3.1. Word Embeddings

Indeed, it has been discussed for quite a long time to use vector representation for words. And recently there has been more and more interest in word embeddings - a technique which maps the words to vectors. Tomas Mikolov's Word2vec algorithm [17] is one driver for this which takes a big volume of text to create high-dimensional representations of words. By this way, it could identify the relationships between words which are not supported by external annotations. Thus, it proves to succeed in getting lots of linguistic regularities with such representation.

In this research, we use word embeddings - Continuous Bag of Words (CBOW) [17]. In the CBOW model, a variety of words illustrate the context for a defined target word which stands for a modification of neural network architecture. For example, with "trèo" (climbed) as the target word, let's use "mèo" (cat) and "trên" (on) as the words to describe the context. Figure 2 describes this model as below:

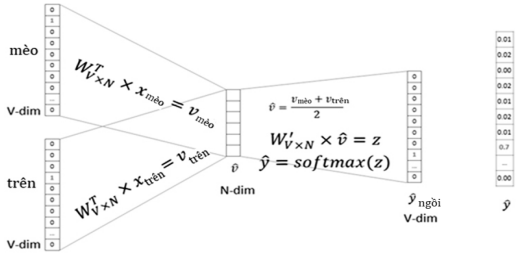

Figure 2: Continuous Bag of Words (CBOW)

3.2. LSTM

Long short-term memory (LSTM) [18] is a modification of the Recurrent neural networks (RNN). Thanks to the feedback loop in its architecture, this model has the ability to store the records of previous outputs.

Figure 3 represents a diagram of a simple LSTM cell. As befits the term deep neural networks, all single cells are combined altogether to shape a huge network. The cell unit represents the memory with five main elements: an input gate *i*, an output gate *o*, a forget gate *f*, hidden state output *h* and a recurring cell state *c*. Given a sequence of vectors (x_1 ,x_2, ..., x_n), σ is the logistic sigmoid function, the hidden state h_t of LSTM at time t is calculated as follows:

$$h_t = o_t * tanh(c_t) \quad (1)$$

$$o_t = tanh(Wx_0x_t + Wh_0h_{t-1} + Wc_0c_t + b_o) \quad (2)$$

$$c_t = f_t * c_{t-1} + i_t * tanh(Wx_cx_t + Wh_ch_{t-1} + b_c) \quad (3)$$

$$f_t = \sigma(Wx_fx_t + Wh_fh_{t-1} + Wc_fc_{t-1} + b_f) \quad (4)$$

$$i_t = \sigma(Wx_ix_t + Wh_ih_{t-1} + Wc_ic_{t-1} + b_i) \quad (5)$$

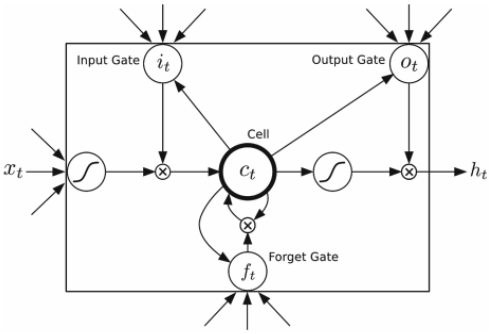

Figure 3: The Long Short Term Memory cell (Source: [19])

3.3. Bidirectional LSTM

Bidirectional LSTM [20] could be seen as a suitable model in many tasks of natural language processing. This model is modified version of LSTM by combining forward and backward LSTMs. The Bi-LSTM model is shown in Figure 4.

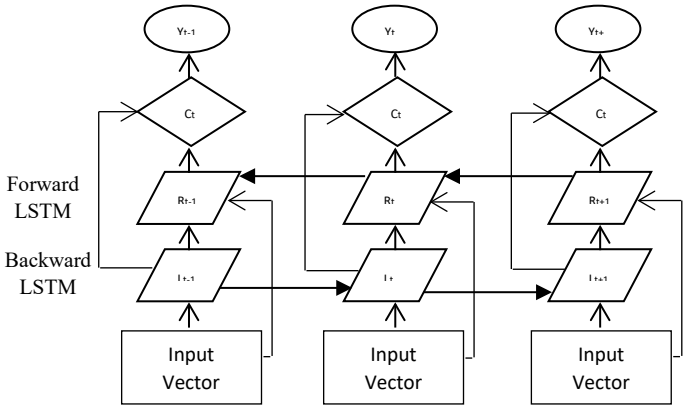

Figure 4: Bidirectional LSTM

Assuming that above equations are abbreviated to

$$h_t = \text{LSTM}(x_t\ h_{t-1}). \quad (6)$$

The forward $\overrightarrow{ht}$ and $\overleftarrow{h}$t the backward LSTM are defined as follows:

$$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}}) \quad (7)$$

$$\overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t-1}}) \quad (8)$$

Where → denotes the forward and ← denotes backward. By taking the forward and backward LSTM as input, the Bidirectional LSTM $\overleftrightarrow{ht}$ at time *t* leads to further classification procedure:

$$\overleftrightarrow{h_t} = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (9)$$

3.4. Diacritics Restoration

We used the Encoder-Decoder LSTM model proposed by Bui [6] for Vietnamese diacritics restoration. The Encoder-Decoder model was introduced by Kyunghyun Cho et al. [14]. This model learns to encode a variable-length sequence into a fixed-length vector representation and to decode a given fixed-length vector representation back into a variable-length sequence. This model is presented in Figure 5.

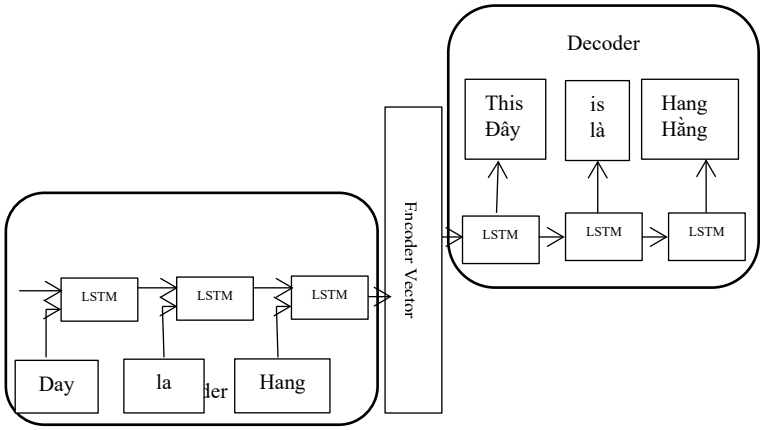

Figure 5: Encoder-Decoder LSTM model

In Figure 5, the input sequence is displayed to the network one encoded character at once. The output produced from this model is a fixed-size vector that shows the internal representation of the input sequence.

3.5. Question Classifier

The output layer is defined as follows based on the text vector studied from the Bidirectional LSTM model:

$$y = (W_d x_t + b_d) \quad (10)$$

where x_t: the text vector studied from the Bidirectional LSTM model; y: the degree of question class of the target text; W_d, b_d: the weight and bias associated with the Bidirectional LSTM model.

The Bidirectional LSTM model is trained by minimizing the mean squared error between the predictive question class and true question class. The loss function is identified as follows:

$$L(X, y) = \frac{1}{2n}\sum_{k=1}^{n}\binom{n}{k}\left\|h(x^i) - y^i\right\|^2 \quad (11)$$

$\mathbf{X} = \{x^1, x^1, x^2, \ldots, x^m\}$: a training set of text matrix

$\mathbf{y} = \{y^1, y^2, \ldots, y^m\}$: a question class ratings set of training sets

4. Experiments

We used dataset in [6] for the first step. This dataset was crawled from the official newspapers website from various categories like sports, world, business and entertainment. The size of dataset is 150 MB. The training dataset is 1500000 sentences and testing dataset is 5000 sentences. We removed all diacritical marks following the rule presented in [6].

To enable a LSTM network to process and remember more effectively, we are going to break dataset down to a lot of n-grams. Also, Vietnamese tone marks for a word can be determined from the small surrounding words in most of the cases; therefore something like a 5-gram or 7-gram model will rather fit for our purpose to some extent. So we have 11 million 5-gram with length varied mostly from 15-25 characters. In addition, as mentioned above, the easiest way to gather a lot of training data to add diacritical marks to Vietnamese text is to get a diacritical text then remove the diacritical marks.

As a machine learning framework, Keras with Tensorflow as a backend has been used. We used same parameters of [6] with 3 LSTM layers, each layer has 256 neurons. We optimized by Adam Optimizer with learning rate 0.002. We used Accuracy score - Accuracy of Diacritics Restoration (ADR) to evaluate the first step as below equation:

$$ADR = \frac{Number\ of\ corrected\ words}{Number\ of\ words} \quad (12)$$

We did experiements with the range of epochs, after about 300 more epochs, or 16 hours we got 97% accuracy and reduced loss to 0.07. Figure 6 shows the Loss of Diacritics Restoration over epochs and Accuracy of Diacritics Restoration per epochs are shown in Figure 7.

Since the used data was extracted from the newspaper websites, it might consist of unpopular words and a couple of errors. Therefore, the diacritic restoration in those cases usually will not achieve highly accurate results.

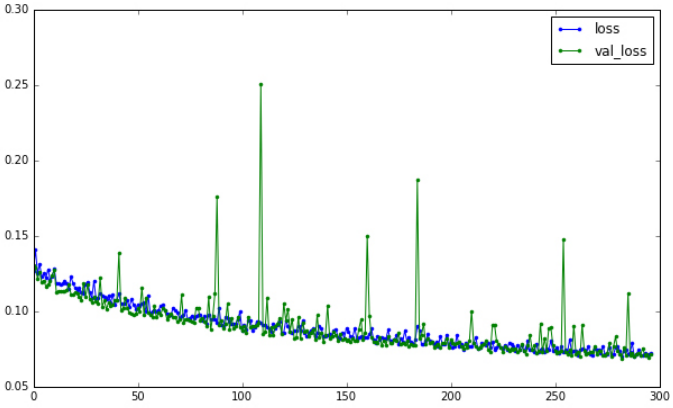

Figure 6: The loss of Diacritics Restoration over epochs

For the second task, we collected question classification dataset from Binh Duong Department of Information and Communications. This data includes 4 areas with 42 types of provincial administrative procedures [21]. The dataset has 560 questions labelled as five primary categories and 270 sub-categories. The main categories and their corresponding number of subcategories are listed in the Table 1. Table 2 shows the details our dataset for training and testing in Question Classification with ratio 8:2.

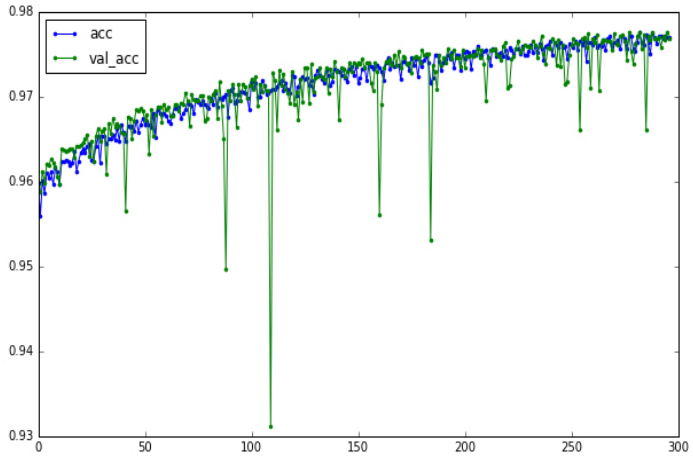

Figure 7: Accuracy of Diacritics Restoration task per epochs

Table 1: Describe of dataset of Question Classification

Categories	Number of Sub-Categories
Journalism	35
Postal	42
Broadcasting	84
Publishing	98
General Information	11

Table 2: Dataset of Question Classification

Name of Dataset	Number of Sentence
The set of questions	570
Sub- Categories	270
Train	456
Test	114

For pre-processing dataset, we used Pyvi (0.0.0.9 - Tran Viet Trung 2016) to tokenize Vietnamese, pre-train Vietnamese word embeddings by streetcodevn (Hung Le 2018). For our training models, we used Tensorflow and Keras libraries. The parameters of our question classification Bi-LSTM model are: Batch size: 500, number of hidden nodes: 128, Drop out: 0.2, epochs: 300, Sigmoid Activate function, Adam Optimization function and Binary cross entropy function. We evaluated performance of question classification by Accuracy of Question Classification (AQC) following below equation:

$$AQC = \frac{\#of\ correct\ predict\ questions}{\#of\ predict\ questions} \quad (13)$$

To evaluate our proposed model, we compared our results with the result of LSTM model separately as illustrated in Table 3. This

result reveals that our proposed model – the Bi-LSTM deep learning approach get the best results. From the result is shown in Table 3, it is obvious that the Bi-LSTM beats the LSTM model; therefore it could be said that the Bi-LSTM is the best-performing score comparing with the LSTM.

Table 3: The results of question classification

Model	AQC
LSTM	92%
Bi-LSTM	**95%**

We made an application by integrating the automated question answering systems into Binh Duong Department of Information and Communications Support System. The name of this application is ICTbot. The ICTbot is used to support Binh Duong Department of Information and Communications in General Information and four management categories: Journalism, Postal, Broadcasting and Publishing. The ICTbot allows users to input new questions and the result is the answers that have the highest accuracy of question classification. Figure 8 shows the ICTbot application.

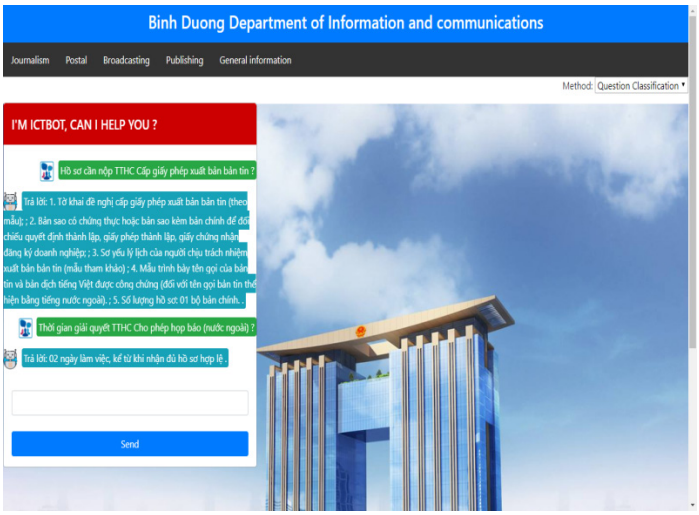

Figure 8: The ICTbot of Binh Duong Department of Information and Communications Support System.

Table 4: The results of question classification in different percentage of Accuracy of Diacritics Restoration

Percentage of Accuracy of Diacritics Restoration	AQC
>= 90%	95%
<= 75 and <90%	70% - 94%
<= 50 and <75%	48% - 69%
<= 25 and <50%	20% -47%
<25%	0-19%

To evaluate the differences in integrating diacritics restoration into question classification, we applied diacritics restoration by percentage into question classification. We counted the number of restoration words in test question dataset and integrated diacritics restoration in test question dataset by percentage of Accuracy of Diacritics Restoration. When percentage of diacritics restoration is equal or larger than 75%, question classification worked well, however when the percentage was so small question classification didn't work well or made a false label of the predict question. The result also depends on the length of question test. Table 4 reveals the result in different percentage of Accuracy of Diacritics Restoration. As a result, it could be seen that integrating diacritics restoration into question classification improves the accuracy of the system.

5. Conclusion

In this paper, we experimented our model by integrating diacritics restoration and question classification into Vietnamese question answering system. The result of the first step will be the input of question classification. It could be said that this is the first research integrating two phases into Vietnamese question answering system. We used the Encoder-Decoder LSTM model for Vietnamese diacritics restoration and the Bidirectional LSTM model for Vietnamese question classification. Our experiments proved that integrating diacritics restoration improves question classification. We also built a useful application based on Question classification – ICTbot and integrated in Binh Duong Department of Information and Communications Support System. In the future, we are looking to apply other deep learning approach and explore more features to improve the results.

Conflict of Interest

I declare no conflict of interest.

References

[1] G. D. Pauw, P. W. Wagacha, and G.-M. de Schryver, "Automatic diacritic restoration for resource-scarce languages," in Proceedings of the Text, Speech and Dialogue, 10th International Conference, TSD, Pilsen,Czech Republic, pp. 170–179, 2007. https://doi.org/10.1007/978-3-540-74628-7_24

[2] M. Simard and A. Deslauriers, "Real-time automatic insertion of accents in French text," Natural Language Engineering , vol. 7, Issue 2, pp. 143–165, 2001. https://doi.org/10.1017/S1351324901002650

[3] Mihalcea, R, "Diacritics restoration: Learning from letters versus learning from words", in Proceedings of CICLing, pp. 339-348. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, 2002. https://doi.org/10.1007/3-540-45715-1_35

[4] Nikola Santic, Jan Snajder, Bojana Dalbelo Basic, "Automatic diacritics restoration in Croatian texts", The Future of Information Sciences, Digital Resources and Knowledge Sharing, pp. 126–130, Springer, 2009.

[5] R. Nelken and S. M. Shieber, "Arabic diacritization using weighted finite-state transducers", in Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, ser. Semitic '05, Stroudsburg, PA, USA, pp. 79–86, 2005. https://doi.org/10.3115/1621787.1621802

[6] Bui Thanh Hung, "Vietnamese Diacritics Restoration Using Deep Learning Approach", in Proceedings of the 10th International Conference on Knowledge and Systems Engineering – KSE, 2018. https://doi.org/10.1109/KSE.2018.8573427

[7] L.N. Pham, T.V. Hong, V.V. Nguyen, "Vietnamese Text Accent Restoration with Statistical Machine Translation", 27th Pacific Asia Conference on Language, Information, and Computation pp.423 — 429, 2013. https://www.aclweb.org/anthology/Y13-1044

[8] N.M.Trung, N.Q.Nhan, N.H.Phuong, "Vietnamese diacritics restoration as sequential tagging", Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), IEEE RIVF International Conference, 2012. https://doi.org/10.1109/KSE.2018.8573427

[9] T.A. Luu, K. Yamamoto, "A Pointwise approach for Vietnamese automatic diacrics restoration", Proceedings of the International Conference on Asian Language Processing-IALP, 2012. https://doi.org/10.1109/IALP.2012.18

[10] Amit Mishra and Sanjay Kumar Jain, "A survey on question answering systems with classification. Journal of King Saud University - Computer and Information Sciences, 28(3), pp 345 – 361, 2016. https://doi.org/10.1016/j.jksuci.2014.10.007

[11] Nguyen Van-Tu and Le Anh-Cuong, "Improving question classification by feature extraction and selection", Indian Journal of Science and Technology, 9(17), 2016. https://doi.org/10.17485/ijst/2016/v9i17/93160

[12] Andreas, J., Rohrbach, M., Darrell, T., and Klein, "Learning to Compose Neural Networks for Question Answering", Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016. https://doi.org/10.18653/v1/N16-1181

[13] Graves, A., "Supervised Sequence Labelling with Recurrent Neural Networks", Studies in Computational Intelligence, vol. 385, pp. 5-13, Springer, 2012.

[14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, 2014. https://doi.org/10.3115/v1/d14-1179

[15] Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria, "Recent Trends in Deep Learning Based Natural Language Processing", IEEE Computational Intelligence Magazin, 2018. https://doi.org/10.1109/MCI.2018.2840738

[16] Yoon Kim, "Convolutional neural networks for sentence classification", Conference on Empirical Methods in Natural Language Processing, pages 1746–1751, 2014. https://doi.org/10.3115/v1/D14-1181

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space", in Proceedings of International Conference on Learning Representations (ICLR-13): Workshop Track, 2013. https://arxiv.org/abs/1301.3781

[18] Hochreiter S., Schmidhuber J., "Long Short Term Memory", Neural Computation 9(8), pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[19] Hong J, Fang M, "Sentiment analysis with deeply learned distributed representations of variable length texts", Technical report, Stanford University, CS224d: Deep Learning for Natural Language Processing, 2015. https://cs224d.stanford.edu/reports/HongJames.pdf

[20] Wang P., Qian Y., Soong F. K., He L., Zhao H., "Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network", 2015. https://arxiv.org/abs/1510.06168

[21] Decision No. 1402/QD-UBND dated May 29, 2018 on the announcement of administrative procedures under the jurisdiction of the Department of Information and Communications/District People's Committees of Binh Duong province.