INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Follow-Up Question Generation using Neural Tensor Network-based Domain Ontology Population in an Interview Coaching System

Ming-Hsiang Su, Chung-Hsien Wu, and Yi Chang

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
{huntfox.su, chunghsienwu, debbiebwe}@gmail.com

Abstract

This study proposes an approach to follow-up question generation based on a populated domain ontology in a conversational interview coaching system. The purpose of this study is to generate the follow-up questions which are more related to the meaning beyond the literal content in the user's answer based on the background knowledge in a populated domain ontology. Firstly, a convolutional neural tensor network (CNTN) was applied for selecting a key sentence from the user answer. Secondly, the neural tensor network (NTN) was used to model the relationship between the subjects and objects in the resource description framework (RDF) triple, defined as (subject, predicate, object), in each predicate from the ConceptNet for domain ontology population. The words in the key sentence were then used to retrieve relevant triples from the domain ontology for filling into the slots in the question templates to generate potential follow-up questions. Finally, the CNTN-based sentence matching model was employed to choose the one most related to the answer sentence as the final follow-up question. This study used 5-fold cross-validation for performance evaluation. The experimental results showed the generation performance in the proposed model was higher than the traditional method. The performance of key sentence selection model achieved 81.94%, and the sentence matching model achieved 92.28%.

Index Terms: Interview coaching, dialogue system, ontology, follow-up question generation, CNTN

1. Introduction

With the advanced progress of artificial intelligence and deep learning technology, spoken dialog systems have been widely applied to various domains, which were generally divided into task-oriented dialog systems [1] and non-task-oriented dialog systems [2]-[3]. There have been many task-oriented dialog systems constructed in the past years, for instance, interview coaching, online shopping, hotel booking, ticket booking, customer service, etc. [4]-[5]. An interview coaching system tries to simulate an interviewer to provide mock interview practice simulation sessions for the users [5]. TARDIS [6] focused on emotional computing and aimed to improve the social skills of young people. Regarding the coaching system with a fixed scenario, MACH [7] analyzed a user's nonverbal behaviors and provided a summary feedback, indicating which nonverbal behaviors were needed to be improved. Although these coaching systems were used to improve user's interview skill, few of them considered follow-up question generation which were more related to the meaning beyond the literal content of the user's answer. In previous studies, most of the interview process in an interview coaching system was pre-defined. Preferably, if the system can ask the follow-up questions based on the background knowledge in a specific domain, the interviewee can practice their interview skills more realistically and effectively.

For follow-up question generation, Moore and Mittal [8] applied templates to generate follow-up questions. They predefined the types of text objects. The users could query and design question templates which considered three types of questions: "ask", "inform" and "recommend". Su et al. [9] adopted a pattern-based sequence-to-sequence (Seq2seq) model for follow-up question generation. Their experimental results showed that their proposed method outperformed the traditional word-based method and achieved a more favorable performance based on a statistical significance test.

The methods for constructing an ontology could be manual [10], semi-automatic [11], or automatic [12]. Khan and Kumar [10] employed the protégé tool to build an ontology. The manually developed ontology had high accuracy but was time-consuming to complete. De Silva and Jayaratne [11] extracted ontologies and modeled them by using the Wikipedia XML database as the source, then manually modified it to meet their requirements. This approach also had high accuracy, but it needed a suitable universal ontology for domain-specific ontology extraction. Lin et al. [12] described an approach to automatically generate an ontology. They used the topic model to generate concepts and used semantic similarity measure to construct the hierarchical structure. The advantage was the ontology could be automatically built by text, but the weakness was the accuracy was lower than that of the manual method. Thus, a semi-automatic approach is suitable for the system in this study to construct a domain-specific ontology with high accuracy.

This study focused on the development and population of a domain ontology for follow-up question generation. The generated follow-up question based on the populated domain ontology was expected to consider the background knowledge beyond the literal content of the user's answer to make the follow-up question more vivid and close to a real interview. The block diagram of the proposed system is shown in Figure 1.

The main contributions of this study are summarized as follows. First, this study used the CNTN-based sentence selection model to select the most appropriate sentence of interviewee's answer as the key sentence. Second, based on the selected key sentence, this study used the key terms with higher term frequency in the collected interview coaching database to construct a domain-specific ontology based on the ConceptNet, which was constructed to give computers access to common-sense knowledge. A triple in the ConceptNet is a set of three entities that codifies a statement about semantic data in the form of subject–predicate–object expressions, denoted as

http://dx.doi.org/10.21437/Interspeech.2019-1300

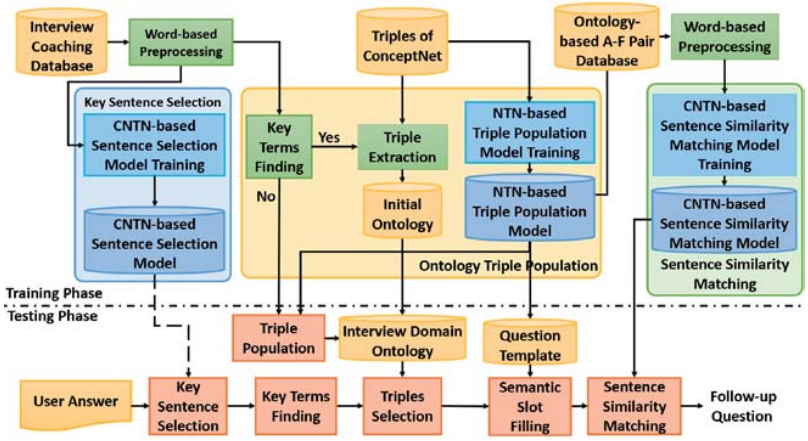

Figure 1: *Block diagram of the proposed follow-up question generation method.*

(subject, predicate, object). If a key term is in the ConceptNet, this study extracts the triples containing the key term for semantic slot filling. If the key term is not in ConceptNet, this study searches the suitable triples containing the key term by using the NTN-based ontology population model to populate the interview domain ontology. Third, this study used the CNTN-based sentence similarity matching model to generate the follow-up question.

2. Database collection

In order to construct an interview coaching system, this study invited twelve participants to collect the interview coaching database. The question types and topics for the interviews were related to the graduate school admission interview. During database collection, every two participants, one serving as the interviewer and the other as the interviewee, had the freedom to complete the interview without using predesigned questions. The interviewee was assigned a random identity to simulate the real situation. In the collected interview coaching database, there were two different questions, namely, ordinary questions and follow-up questions. Ordinary questions were the questions not related to the previous question or interviewee's previous response, while the follow-up questions were asked based on the interviewee's previous response to elaborate the initial response. Finally, 260 dialogs with 1,754 ordinary questions and 1,262 follow-up questions were collected to form the interview coaching database, as shown in Table 1.

Table 1: *Details of the Interview Coaching Database.*

	Total
Number of turns	3,016
Number of ordinary/follow-up questions	1,754/1,262
Average number of turns	10.7
Average number of sentences in each answer	3.84
Average time of interview	20

Besides, this study collected the RDF triples from the ConceptNet which included 362,414 triples and 19 predicates. For constructing a domain ontology, this study used the key terms (6,070 words) with high term frequency in the interview coaching database to extract the triples from the ConceptNet to form an initial domain ontology. After domain ontology extraction, this study found that there were 26 key terms not in the ConceptNet. These key terms were needed to be populated to the initial ontology. This study also invited twelve participants to generate the follow-up questions based on the interviewee's answers and the triples with the subjects or objects included in the answer. These follow-up questions based on the interviewee answers and triples formed an ontology-based answer-follow-up question (A-F) pair database. Finally, this study collected 6,943 A-F pairs.

3. System framework

3.1. Key sentence selection

As the interviewee's response generally contains many sentences, it is challenging to find the most appropriate one as the target sentence for follow-up question generation. This study used the CNTN-based [13] sentence selection model to solve the problem. The CNTN is composed of a convolutional neural network (CNN) [14] and a neural tensor network (NTN) [15], as shown in Figure 2. The CNN is used to encode the sentences of the question and the response, and the NTN is used to learn the relationship between the question and the response sentences.

Given a sentence s, this study used Word2Vec continuous bag-of-words (CBOW) algorithm [16] to obtain the word embedding vector $\mathbf{w}_i \in \mathbb{R}^{n_w}$ for each word w in sentence s. Then the word vector $\mathbf{w}_i$ was used to obtain the input matrix $\mathbf{s} \in \mathcal{R}^{n_w \times l_s}$, where l_s denotes the sentence length. Next, a convolutional layer was obtained by convolving a matrix of weights $\mathbf{m} \in \mathcal{R}^{n \times m}$ with the matrix of activations at the layer below, where m was the filter width. Given a value k and a row vector $\mathbf{p} \in \mathcal{R}^{p}$, this study used k-max pooling to select the subsequence $\mathbf{p}^{p}_{max}$ of the k highest values of $\mathbf{p}$. The k-max pooling operation made it possible to pool the k most active features in $\mathbf{p}$. The final output of CNN was a vector $\mathbf{v}_s \in \mathcal{R}^{n_s}$, which represented the embedding of the input sentence s. Given a sentence of interviewee's response q and a sequence r where r was formed by the interviewee's response in sequence, this study modeled $\mathbf{v}_q$ and $\mathbf{v}_r$ by using the CNN. Then the tensor layer calculated the relevance score of a question-response pair by (1).

$$\text{score}(q,r) = \mathbf{u}^T f(\mathbf{v}_q^T \mathbf{M}^{[1:a]} \mathbf{v}_r + \mathbf{V}\begin{bmatrix}\mathbf{v}_q \\ \mathbf{v}_r\end{bmatrix} + \mathbf{b}) \quad (1)$$

where f was a standard nonlinearity applied element-wise, $\mathbf{V} \in \mathcal{R}^{a \times 2n_s}$, $\mathbf{b} \in \mathcal{R}^{a}$, $\mathbf{u} \in \mathcal{R}^{a}$, $\mathbf{M}^{[1:a]} \in \mathcal{R}^{n_s \times n_s \times a}$ was a tensor and the bilinear tensor product $\mathbf{V}_q^T \mathbf{M}^{[1:a]} \mathbf{v}_r$ resulted in a vector $h \in \mathcal{R}^{a}$, where each entry was computed by one slice $i = 1, \ldots, a$ of

the tensor $h_i = \mathbf{V}_q^T \mathbf{M}^i \mathbf{v}_r$. This study selected the sentence with the highest maching score from the interviewee's response as the key sentence by using the CNTN-based target sentence selection model. The selected sentence was then used for ontology population and follow-up question generation.

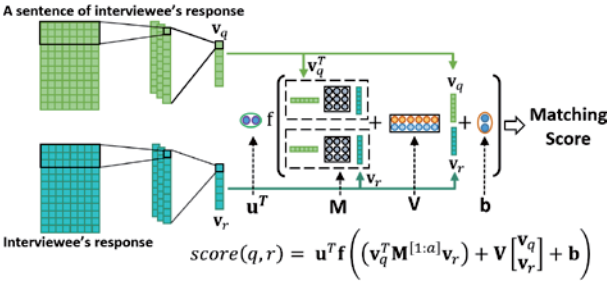

Figure 2: *CNTN-based target sentence selection model.*

3.2. Ontology extraction and population

To build a reliable domain ontology, it is necessary to rely on the existing universal ontology for its high accuracy. Firstly, this study used the follow-up questions in the interview coaching database to find the words with high term frequency (TF) as the key terms. In this study, stop words were removed. In this study, the follow-up questions were regarded as the documents in which 104 words were extracted as the key terms. Besides the 104 words, this study additionally considered 5,966 Chinese frequently-used words as the key terms. Finally, 6,070 words were defined as the key terms in this study. Then, the key terms were used to extract the triples from the ConceptNet to form an initial ontology. There were 26 words which were not found in the ConceptNet. Thus, these 26 words were considered as the terms needed to be populated to the initial ontology.

For ontology population, this study trained 19 NTN-based triple population models for 19 predicates in the ConceptNet, as shown in Figure 3. For training 19 NTN-based triple population models, the subject and object of the triples were transformed into word vectors as the inputs of the NTN model. Then every two words of the key terms along with a predicate were verified to check if they can form a triple using the trained NTN models.

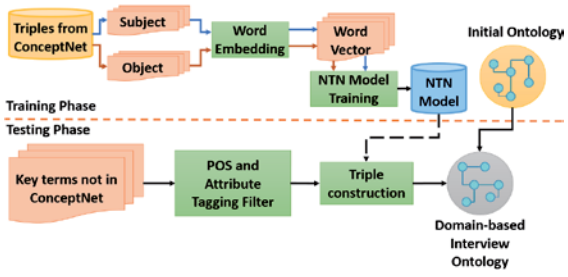

Figure 3: *Ontology triple population process.*

To construct reasonable triples, it is needed to select suitable subjects and objects for ontology population. This study used the part of speech (POS) tag and attribute based on the E-HowNet [17] to solve this issue. According to the statistical distribution for the subjects and objects of each predicate, this study only considered the words that matched the top three POS tags and attributes in each predicate. For instance, the top three POS tags of the subjects of "AtLocation" predicate were Nab (countable individual nouns), Naa (uncountable entity nouns) and Nad (uncountable non-entity nouns) and the top three attributes were telic, predication and location. If the POS tag of a new word was Nab or the attribute of a new word was location, the new word was tested as the subject for the "AtLocation" predicate model. The complexity of the input words was reduced through this approach and more reasonable triples were obtained.

3.3. Sentence similarity matching

After key sentence selection, this study selected the key sentence for follow-up question generation. This study also extracted all relevant triples with the key terms of the sentence from the domain ontology. Then this study used relevant triples to generate follow-up questions based on the question templates and used the CNTN-based sentence matching model to choose the most suitable follow-up question, as shown in Figure 4. For example, the key sentence was "我大三的專題研究是影像處理相關 (My junior research project is related to image processing)." Some relevant triples of the key terms of the sentence are <大三(junior), NotDesires, 二一(quit school)>, <專題(project), MadeOf, 論文(thesis)>, <專題(project), MadeOf, 知識(knowledge)>. All relevant triples were separately filled into the follow-up question templates of the predicates, such as "你的<sub>有參考別的<obj>完成嗎? (Is your <sub> done with reference to other <obj>?)". Finally, this study used the CNTN-based sentence similarity matching model to calculate the matching score between the key sentence and the generated follow-up questions. The generated follow-up question with the highest matching score was selected as the most suitable follow-up question, such as "你的專題有參考別的知識完成嗎? (Is your research project done with reference to other knowledge?)".

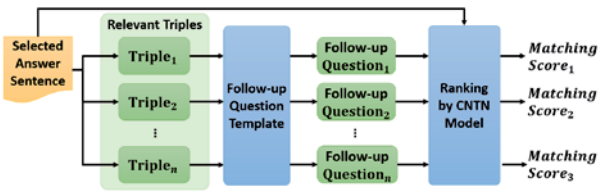

Figure 4: *The Processing of follow-up question generation.*

4. Experimental results and discussion

4.1. Key sentence selection

In this study, the CNTN model was applied to model the relation between each answer sentence and the entire answer turn. The training data was tagged as relevant and irrelevant by five participants. To ensure the consistency, relevant data were needed to be tagged by at least three people. Through this tagging approach, totally 5,112 training data were collected, including 2,814 relevant data and 2,298 irrelevant data. In this study, five-fold cross validation was applied to evaluate the effect of different methods. First, this study evaluated the effect of different CNN filter numbers and NTN tensor dimensions of the CNTN model to select the parameters with best performance. According to the experimental results, the highest accuracy of the CNTN model was 81.94% when the filter number was 32 and the tensor dimension was 3, as shown in Figure 5. Then this study evaluated the performance of CNTN model and the traditional TFIDF with cosine similarity method.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. In this study, TF-IDF values were used to form the sentence representation vector, then cosine similarity was calculated to measure the value between the two sentence vectors. The experimental result showed that the accuracy of CNTN was 81.94% and the accuracy of TFIDF with cosine similarity method was 56.06%. The CNTN outperformed the method using TFIDF and cosine similarity.

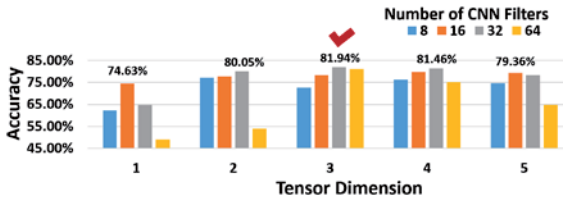

Figure 5: *Effect of CNN filter number and NTN tensor dimension.*

4.2. Ontology triple population

For triple population, this study evaluated 19 NTN models for 19 predicates of the ConceptNet. First, five-fold cross validation was applied to evaluate the performance of 19 NTN models with different tensor dimensions. The experimental results are shown in Figure 6.

Predicate	TM*	Accuracy	Predicate	TM*	Accuracy
AtLocation	4	83.76%	CapableOf	1	67.97%
Causes	1	68.19%	CausesDesire	1	69.36%
DerivedFrom	5	80.50%	Desires	1	74.49%
HasA	5	84.16%	HasFirstSubevent	1	74.06%
HasProperty	3	74.68%	HasSubevent	1	68.52%
IsA	4	79.27%	MadeOf	5	82.04%
MotivatedByGoal	1	71.13%	NotDesires	2	78.19%
PartOf	5	76.41%	RelatedTo	5	82.30%
SymbolOf	5	70.95%	Synonym	5	83.34%
UsedFor	5	85.13%	**TM***: Tensor dimension		

Figure 6: *Effect of NTN tensor dimension for19 predicates.*

Then, this study compared the NTN model with the linear model and bilinear model, as shown in Figure 7. The average accuracy of NTN model was 76.59%, the average accuracy of linear model was 74.95% and the average accuracy of bilinear model was 76.03%. The experimental results showed that the performance of the NTN model was better than the linear model and bilinear model. However, the linear model performed the best on the predicates "CapableOf" and "DerivedFrom", and the bilinear model performed the best on the predicates "Causes", "Desires" and "MotivatedByGoal". After training the NTN models, this study applied the words with high TF in the interview coaching database for ontology population. The experiments were conducted on the data of three types. The first type was the triples consisting of 150 selected words; the second type was the triples consisting of 100 selected words; the third type was the triples consisting of 50 selected words. The words used in the three types were selected based on the term frequency of the words. In these three types, there were new terms which did not exist in the ConceptNet and were needed to be populated to the ontology. Then every two words in each type selected randomly were fed into the NTN models for ontology population. Finally, subjective evaluation was applied to evaluate the accuracy of the populated triples. In the first type, the number of populated triples was 141, in which only 93 triples were suitable for population (accuracy was 65.96%). In the second type, the number of populated triples was 95 and only 67 triples were suitable for population (accuracy was 81.05%). In the third type, the number of populated triples was 12 which were all suitable for population (accuracy is 100%). The experimental result showed that the proposed method was suitable for automated triple population when a small number of words was needed to populate.

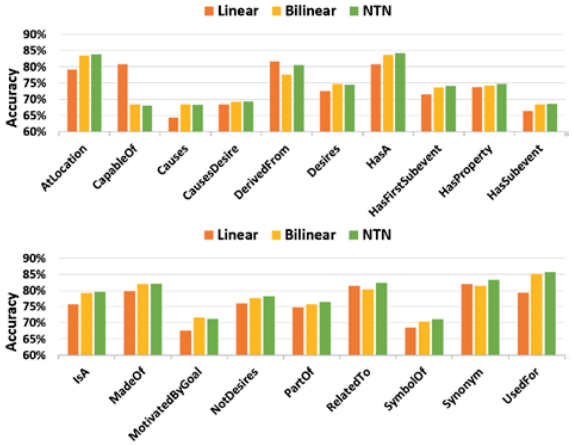

Figure 7: *The performance of NTN compared with Linear and Bilinear model.*

4.3. Sentence similarity matching

After all relevant triples of the key terms without stop words were found in the sentence, this study used the relevant triples to generate follow-up questions based on the question templates and used the CNTN model to choose the one most related to the user's answer sentence as the follow-up question. Five-fold cross validation were applied to find the best setting for the CNTN model. The experimental result showed that the CNTN model had the best performance when the CNN filter was 16 and the tensor dimension was 4. This study also compared the CNTN model and the traditional TF-IDF with cosine similarity model. The accuracy of CNTN model was 92.28% and the accuracy of TF-IDF with cosine similarity model was 51.59%. The performance of CNTN model outperformed the traditional model.

5. Conclusions

This study proposes an approach to follow-up question generation based on a domain ontology in a conversational interview coaching system. Firstly, a CNTN was applied for selecting a key sentence. Secondly, the NTN was used to model the relationship between the subjects and objects in a RDF triple for each predicate in the ConceptNet, and the NTN model was also used to populate the ontology. After extracting the words in the key sentence to query the ontology for relevant triple retrieval, these retrieved relevant triples were filled into the slots in the question templates to output the potential follow-up questions. Finally, this study employed the CNTN-based sentence matching model to choose the one most related to the answer sentence as the final follow-up question. This study applied 5-fold cross validation for evaluation. The experimental results showed that the proposed methods outperformed the traditional methods. In the future, this study will collect more interview data on different domains to construct the ontology in different domains. Besides, this study hopes to record the user's learning process to improve the system performance.

6. References

[1] Y. Mu, and Y. Yin, "Task-oriented spoken dialogue system for humanoid robot," *in 2010 International Conference on Multimedia Technology, October 29-31, Ningbo, China, Proceedings*, 2010, pp. 1-4.

[2] M. Koshinda, M. Inaba, and K. Takahashi, "Machine-learned ranking based non-task-oriented dialogue agent using twitter data," in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), December 6-9, Singapore, Singapore, Proceedings*, 2015, pp. 5-8.

[3] M.-H. Su, C.-H. Wu, K. Y. Huang, and W.-H. Lin, "Response Selection and Automatic Message-Response Expansion in Retrieval-Based QA Systems using Semantic Dependency Pair Model," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 18, no. 1, pp. 3:1-3:23, 2018.

[4] X. Liu, and W. Zhao, "Buddy: A Virtual Life Coaching System for Children and Adolescents with High Functioning Autism," in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), November 6-10, Orlando, FL, USA, Proceedings*, 2017, pp. 293-298.

[5] M.-H. Su, C.-H. Wu, K.-Y. Huang, and C.-K. Chen, "Attention-Based Dialog State Tracking for Conversational Interview Coaching," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 15-20, Calgary, AB, Canada, Proceedings*, 2018, pp. 6144-6148.

[6] H. Jones and N. Sabouret, "TARDIS-A simulation platform with an affective virtual recruiter for job interviews," in *Intelligent Digital Games for Empowerment and Inclusion (IDGEI), May 2-4, Chania, Crete, Greece, Proceedings*, 2013, pp. 1-8.

[7] M. E. Hoque, M. Courgeon, J. C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *the 2013 ACM international joint conference on Pervasive and ubiquitous computing, September 8-12, Zurich, Switzerland*, *Proceedings*, 2013, pp. 697-706.

[8] J. D. Moore, and V. O. Mittal, "Dynamically generated follow-up questions," *Computer*, vol. 29, no. 7, pp. 75-86, 1996.

[9] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, and H.-H. Huang, "Follow-up Question Generation Using Pattern-based Seq2seq with a Small Corpus for Interview Coaching," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018, pp. 1006-1010.

[10] J. A. Khan, and S. Kumar, "Deep analysis for development of RDF, RDFS and OWL ontologies with protégé," in *3rd International Conference on Reliability, Infocom Technologies and Optimization, October 8-10, Noida, India, Proceedings*, 2014, pp. 1-6.

[11] L. N. De Silva, and L. Jayaratne, "WikiOnto: A system for semi-automatic extraction and modeling of ontologies using Wikipedia XML corpus," in *2009 IEEE International Conference on Semantic Computing, September 14-16, Berkeley, CA, USA, Proceedings*, 2009, pp. 571-576.

[12] Z. Lin, R. Lu, Y. Xiong, and Y. Zhu, "Learning ontology automatically using topic model," in *2012 International Conference on Biomedical Engineering and Biotechnology, May 28-30, Macao, China, Proceedings*, 2012, pp. 360-363.

[13] X. Qiu, and X. Huang, "Convolutional Neural Tensor Network Architecture for Community-Based Question Answering," in *the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), July 25-31, Buenos Aires, Argentina, Proceedings,* 2015, pp. 1305-1311.

[14] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1-20, 2016.

[15] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in Neural Information Processing Systems, December 5-10, Lake Tahoe, Nevada, USA, Proceedings*, 2013, pp. 926-934.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems, December 5-10, Lake Tahoe, Nevada, USA, Proceedings*, 2013, pp. 3111-3119.

[17] M.-H. Su, C.-H. Wu, K.-Y. Huang, and Q.-B. Hong, "LSTM-based text emotion recognition using semantic and emotional word vectors," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), May 20-22, Beijing, China, Proceedings*, pp. 1-6.