ICIC Express Letters
Volume **14**, Number **1**, January **2020**

ICIC International ©2020 ISSN 1881-803X
pp. 67–73

AUTO-GROWING KNOWLEDGE GRAPH-BASED INTELLIGENT CHATBOT USING BERT

SoYeop Yoo and OkRan Jeong*

Department of Software
Gachon University
1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do 13120, Korea
bbusso@gc.gachon.ac.kr; *Corresponding author: orjeong@gachon.ac.kr

Received July 2019; accepted October 2019

Abstract. *It is very important to get computers to understand human common sense. A knowledge graph that links words based on relationships is an important skill that allows computers to learn common sense easily. However, knowledge graphs, devised by many existing studies, consist only of specific languages or fields and have limitations that cannot treat neologisms. In this paper, we propose a chatbot system that collects and analyzes data in real-time to build an automatically scalable knowledge graph and utilizes it as the base data. In particular, the fine-tuned BERT-based model for relation extraction is to be applied to auto-growing graphs to improve performance. By building a chatbot where human common sense is learned by using auto-growing knowledge graph, it verifies the availability and performance of knowledge graph.*

Keywords: Knowledge graph, Chatbot, BERT, Relation extraction

1. **Introduction.** Today, interest in artificial intelligence technology is rising and various studies are underway as the improved performance of the hardware, which enables the rapid processing of countless accumulated big data and complex computations. In particular, the fourth industrial revolution centers on the convergence of intelligence technologies in information technology and artificial intelligence software such as big data and IoT. Many studies are focused on artificial intelligence technology to analyze or predict collected data, and more recently, artificial intelligence technology is being integrated with other areas and technologies [1-3].

A chatbot is the most easily accessible system for convergence technology with artificial intelligence technology. Research is being carried out to enhance the performance of chatbot by applying artificial intelligence technology to natural language processing, recognition, and creation in order to communicate with humans. Among the various artificial intelligence technologies required for the chatbot [1-3], making it possible for computers to understand human common sense is a very important technology. For humans, language is a kind of mutual protocol that learns socially and culturally naturally, but computers need separate learning to understand human common sense. Of the many technologies that enable this, knowledge graphs can be used as key underlying data in the chatbot system as they link words and words into relationships and represent them as graphs [4,5].

Knowledge graphs are key skills for computers to understand the meaning of words and the relationship between words, just like humans. In order to understand sentences more like humans, we need a knowledge graph technique that links words to words. Various studies are underway on knowledge graphs such as WordNet, YAGO, Probase, and ConceptNet [4-9].

DOI: 10.24507/icicel.14.01.67

WordNet [6] is a database of words for English and is classified as a noun, verb, adjective, etc., based on the meaning of words. It classifies as synsets and links to each other in conceptual meaning and relation. It consists of about 207,000 word-meaning pairs. YAGO [7] is a knowledge graph built on Wikipedia, WordNet, and GeoNames data. There are more than 10 million people connected. Probase [8] is a knowledge graph built by Microsoft that collects data from approximately 1.6 billion web pages. It is made up of about 2 million objects and more than 200 million pairs. ConceptNet [9] is an open-source knowledge graph that has a large amount of data on 10 key languages, including English and French, and a small amount of data on about 68 languages, including Korean. More than 8 million people are connected based on 40 relationships with each other.

Traditional knowledge graphs, however, have limitations that are focused only on specific languages or fields, or that cannot respond to new words.

In addition, various knowledge graphs are being studied, but existing knowledge graphs still require a lot of human intervention, including the expansion of knowledge graphs at a time more than a certain time after the data is accumulated. Also, because data is used based on specific languages and disciplines, language dependencies exist and are difficult to respond to new words.

In this paper, we are going to apply knowledge graphs that can help us understand human common sense to chatbots that talk with humans so that we can learn general knowledge, common sense from computers. Polaris [10], real-time big data analysis and prediction system, processes and analyzes data efficiently. Polaris uses a graph of automatically extended knowledge, PolarisX, as the underlying data, with real-time data collection, event detection, path analysis, emotion analysis, and predictability. PolarisX automatically expands knowledge graphs by collecting and analyzing new data sources in real-time. In particular, using Google's BERT [11], which shows high performance in the recent NLP field, we are looking to apply a relationship extraction model to building a higher-performance knowledge graph. Based on this PolarisX, we designed and implemented PolarisX-bot, a chatbot that shows the relationship to related knowledge.

The rest of this paper is organized as follows. Section 2 describes the proposed system, PolarisX-bot, an automatically extended knowledge graph-based chatbot. We verify our proposed system and show the results in Section 3. At last, we conclude and discuss future work in Section 4.

2. The Proposed System.

PolarisX-bot, an automatic extended knowledge graph-based intelligent chatbot proposed in this paper, was designed as shown in Figure 1. The system consists of a knowledge graph layer and a chatbot layer. Auto-extend knowledge graphs expand existing graphs by expanding data sources through crawlers and extracting new relational knowledge. Apply the ever-expanding knowledge graph as the basis of chatbot so that human common sense can be understood.

2.1. Auto-growing knowledge graph using BERT.

The knowledge graph links the relationships of words, sentences, and so on to represent knowledge graphically. It contains a relationship of knowledge, which can be useful for computers to learn human common sense. However, many existing knowledge graphs are concentrated only in certain languages, such as English and French, and are based on previously collected data such as Wikipedia, so they do not respond to various languages or creed. It is necessary to improve these limitations and to establish knowledge graphs that are responsive to new words and have no language dependencies.

Auto-growing knowledge graphs expand the data sources to be analyzed to improve the limits of existing knowledge graphs. We collect data from news, social media and others in real-time and analyze the collected data to extract the relationship between words (entities). The extracted words (entities) – relationship pair, automatically expands the

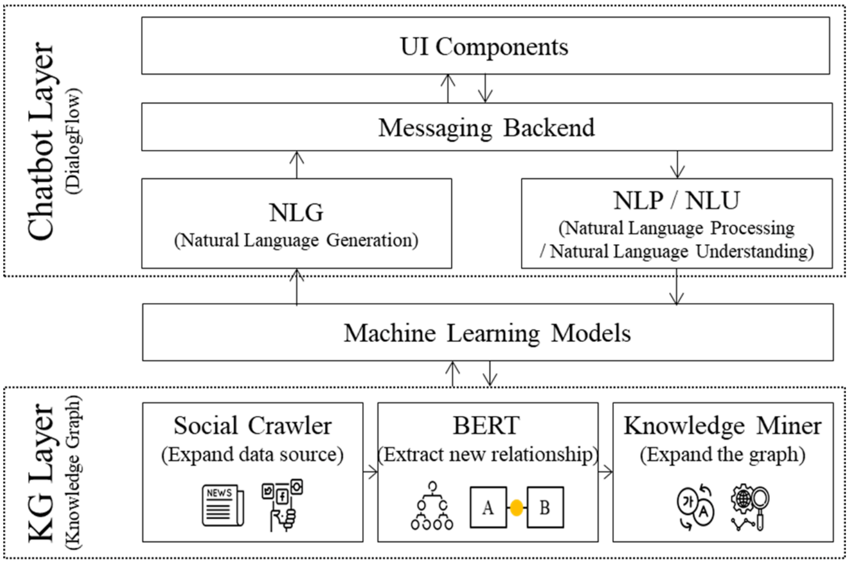


FIGURE 1. System architecture of the proposed chatbot

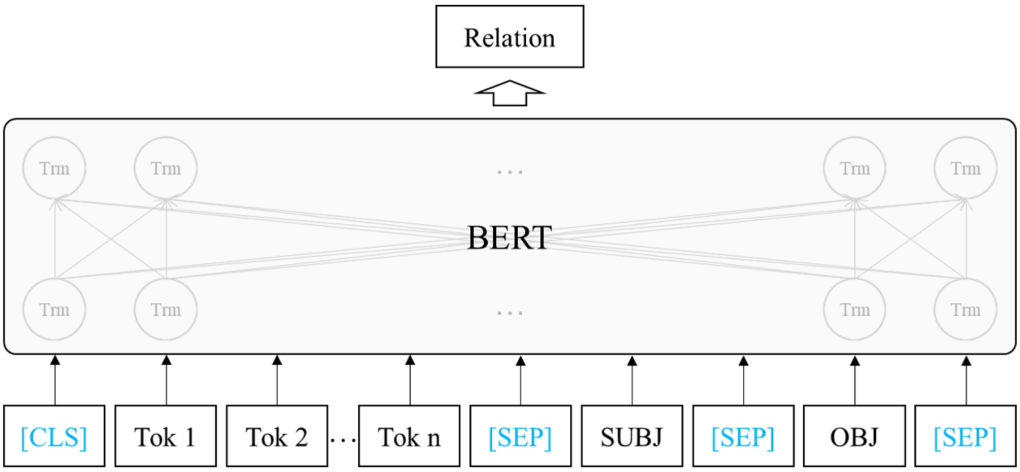


FIGURE 2. Relation extraction using BERT

graph by checking whether it exists in the existing knowledge graph and adding it if it does not exist.

We use the BERT [11] model to extract new relationships from the construction of the automatic expansion knowledge graph. BERT is a pre-trained language model released by Google that has taken up state-of-the-art in 11 tasks in the NLP field. BERT, which learns existing data in advance and is released as a general language model, can be fine-tuned using learning data according to the task you want to perform. We use BERT to tune models to extract relationships between keywords (entities).

We use the relationship-based data, TACRED [12], so that the BERT model can be used in the automatically extended knowledge graph. TACRED data is a relation extraction data set made from news or web text from the corpus of the TAC Knowledge Base Population Challenge. Examples cover 41 types of relationships and are labeled 'no_relation' if they are undefined.

The extraction of relationships requires locating the relationship that represents the relationship with two identities. Figure 2 shows the inputs and outputs of the proposed relationship extraction model. For the learning of the relationship extraction model, when inputting a sentence into the BERT model as input, the subject (SUBJ) and object (OBJ) provided by TACRED are concatenated after the whole sentence. [CLS] refers to the start of the input and [SEP] refers to the separator. The whole sentence is tokenized by the BERT tokenizer and shown as 'Tok 1, Tok 2, . . . , Tok n' in Figure 2. The concatenated sentence goes into the BERT model as an input example. The final output, i.e., label, is a relation. This learned and verified relation extraction model is used to extract relationships in new sentences.

We collect social media Twitter and news data in real-time to extract the most reported keywords. The corresponding sentence with the extracted keywords on Twitter and News data goes as an input of the BERT-based relation extraction model. We extract the relationship between keywords and other words and then make a pair with the extracted entities and relationships such as {*keyword, relation, word*}. We extend the graph with the created pair by mapping it with the existing knowledge graph.

PolarisX extends based on ConceptNet [9], an open-source knowledge graph among existing knowledge graphs. Data sources to gain new knowledge of relationships are based on Twitter and News. The AsterixDB [13], an open-source big data management system, enables real-time collection and storage of Twitter data by entering simple settings and queries through the *FeedAdapter* feature. It can also be used as a database system for PolarisX as various indexing techniques can improve efficiency in processing big data.

We collect key keywords from Twitter and News data collected in real-time and extract the relationships by extracting sentences that contain keywords. We fine-tune the BERT model based on TACRED data for relation extraction. Then we create a new relationship knowledge pair by extracting the linked relationship knowledge between keywords and other words based on the model. Knowledge pair is connected in the form of {*keywords, relationships, words*}. Finally, continue to build scalable knowledge graphs through comparison mapping with existing knowledge graphs.

2.2. Intelligent chatbot using auto-growing knowledge graph.

Chatbot, which carries out dialogue with humans, needs to deal with human natural language to understand its intentions and give an answer that corresponds to its intention. The chatbot system consists mainly of modules to recognize natural language such as human speech and text, modules to grasp its intentions, and modules to make answer contents into natural language again according to its intention. In fact, a number of modules and models are needed to build a chatbot system from start to finish. In this paper, Google's interactive chatbot service, DialogFlow[1], is used to implement the chatbot system easily.

Google's DialogFlow is a service that helps build the chatbot you need through several settings. We use PolarisX, an auto-growing knowledge graph, as the underlying data to build a PolarisX-bot. We use natural language recognition and processing module from DialogFlow to analyze sentences entered by the user, search for corresponding keywords in PolarisX, and show the result to the user. Because there can be various results for one keyword and relationship, it is more intuitive to show knowledge by graphically showing the top five results, rather than simply showing the results in sentences.

The PolarisX-bot is built using the auto-extended knowledge graph PolarisX and DialogFlow. Because the general knowledge and common sense of humans can learn chatbot through the knowledge graph and thus communicate accordingly, the utilization of auto-growing knowledge graph can be verified through the intelligent chatbot.

3. Experiments.

We establish knowledge graphs that automatically expand using BERT models, and implement intelligent chatbot that can exchange conversations based on them. To verify the proposed system, the accuracy of the BERT-based automatic expansion knowledge graph is measured, and the implementation results of an intelligent chatbot are shown.

[1] https://dialogflow.com/

3.1. **Dataset.** For the experiment of proposed systems, two main classes of data are used. The first is social media and news data that BERT-based auto-extended knowledge graphs collect in real-time to keep expanding. The second data is for learning the deep learning model, which allows new relationships to be extracted from the collected data.

Table 1 summarizes the datasets used in the experiment. Twitter and News are used to expand data sources on knowledge graphs. Actually, data can be collected in real-time and the knowledge graph can be automatically expanded, but we use data from November 2018 for experiments. Twitter data is collected in real-time using Apache AsterixDB's *FeedAdapter* function, while News data is collected through NewsAPI[2].

Table 1. Dataset

Data	Size	Collection method
Twitter	About 15 million tweets	FeedAdapter (AsterixDB)
News	About 102,000 articles	News API
TACRED	About 106,000 sentences	Linguistic Data Consortium

TACRED data are training data of deep learning models for extracting new relationships from the new data such as Twitter and News data. It is also used to verify the relation extraction model of the knowledge graph through experiments. The TACRED data is available through the Linguistic Data Consortium, which consists of a total of 106,264 sentences. A set of relationship extraction data created from a corpus used in the TAC Knowledge Base Population Challenge that contains a total of 42 relationships, including 'no_relation'.

3.2. **Results on auto-growing knowledge graph.** It is very important that the proposed auto-growing knowledge graph finds and graphs new relationships from new data. We train the pre-trained BERT model using TACRED dataset to fine-tune for the task of relation extraction and build a deep learning model for relation extraction. We use the TPU environment in Google colab[3] to build a model.

BERT has different pre-trained models depending on case-sensitive status, a number of layers, and a number of hidden units. We use the model bert_cased_L-12_H-768_A-12 that is case-sensitive, consisting of 12 layers and 768 hidden units, in experiments for verification of the auto-growing knowledge graph. We also use a TACRED dataset to train a model for use in the knowledge graph. We experiment with the TACRED dataset by separating the train, dev, and test set at about 65%, 20%, and 15%, respectively.

Table 2 shows the experimental results of the relationship extraction model. The accuracy was about 0.75 for the dev set and 0.79 for the test set. Relation extraction is more complex than other existing NLP tasks. In relation extraction task, we should extract the exact subject and object from the sentence first, then extract relation between them. Actually there are two steps in relation extraction. It is expected that the application of the BERT-Large model with 24 layers with a relatively larger model will perform better because it showed good results when performed with the BERT-Base model.

Table 2. Experiment results on BERT-based relation extraction model

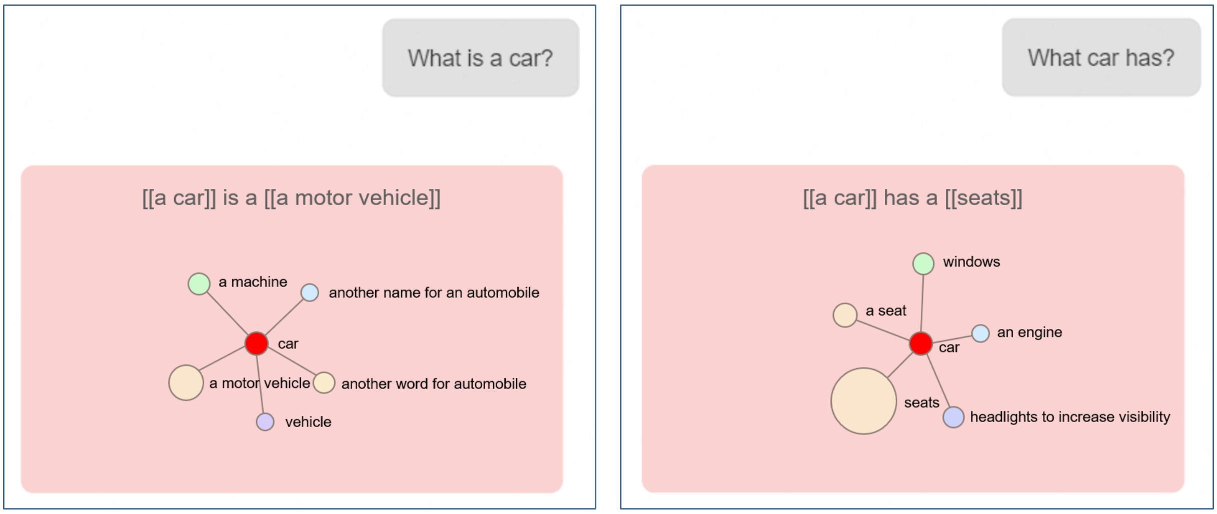
BERT model	Dataset	Evaluation set	Accuracy	Loss
bert_cased_L-12_H-768_A-12	TACRED	dev set	0.7528	1.2011
		test set	0.7885	1.1412

[2] https://newsapi.org/

[3] https://colab.research.google.com/

3.3. **Results on intelligent chatbot.** In this paper, ConceptNet, which is utilized as a base knowledge graph of the automatic expansion knowledge graph, has 40 relationships including *IsA*, *HasA*, *PartOf*, and so on. If there is an *IsA* relationship between the *object* A and the *object* B, it indicates that A is a subclass or example of B. Therefore, all A's belong to B. The *HasA* relationship means that B belongs to A. Normally *HasA* is also the opposite of *PartOf* relations [9].

Figure 3 shows an example of a conversation using a built-in automatic expansion knowledge graph-based chatbot. It shows the answer to the question 'What is a car?' and 'What car has?' based on the auto-growing knowledge graph, respectively. The left figure shows the results related to the car and *IsA* relationship as a motor vehicle, machine, vehicle, and so on with the different sizes of nodes depending on weight. The answer with the highest weight, the motor vehicle, is also answered with a sentence. The right figure shows the words seats (seats), windows (windows), engines, and so on as the results in the relationship between the car and *HasA*.


What is a car?

[[a car]] is a [[a motor vehicle]]

What car has?

[[a car]] has a [[seats]]

a machine
another name for an automobile
car
a motor vehicle
another word for automobile
vehicle
windows
a seat
an engine
car
seats
headlights to increase visibility

FIGURE 3. Example of the intelligent chatbot

The auto-growing knowledge graph-based intelligent chatbot shows that it is based on the knowledge graph and presents the results in graph form to show users multiple answers rather than one answer. By visualizing results in the graph as well as through sentences, it enables more intuitive results to be identified.

4. **Conclusion.** Now that artificial intelligence technology is actively being studied and utilized, the technology to teach computers human common sense is a very important technology. The relationship of words that reflect human common sense can be graphed out and used as the underlying data, making it easy for computers to learn human common sense.

In this paper, we propose an intelligent chatbot that builds and applies BERT-based auto-growing knowledge graph. It was verified through experiments and implementation that auto-growing knowledge graph does not have language dependencies, can respond to new words, and can be used in various ways as opposed to existing knowledge graphs.

In future work, we would improve the way to extract relationships. The fine-tuned BERT model shows the good result on the TACRED dataset, but it still has a limitation that the proposed model can extract relations only in the labels of the dataset. To improve this problem, we could use distant supervision to figure out new relationships not in the dataset. Also, we would expand our intelligent chatbot as a framework for conversational AI.

Acknowledgment. This research was supported by Basic Science Research Program through the NRF (National Research Foundation of Korea), and the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute for Information & communications Technology Promotion), and Copyright Technology R&D through the MCST (Ministry of Culture, Sports and Tourism), Korea and the KCC (Korea Copyright Commission) (Nos. NRF 2019R1A2C1008412, 2015-0-00932, 2019-CONTEXT-9500).

REFERENCES

[1] G. Ji, S. He, L. Xu, K. Liu and J. Zhao, Knowledge graph embedding via dynamic mapping matrix, *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp.687-696, 2015.

[2] Y. Lin, Z. Liu, M. Sun, Y. Liu and X. Zhu, Learning entity and relation embeddings for knowledge graph completion, *The 29th AAAI Conference on Artificial Intelligence*, 2015.

[3] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation method, *Semantic Web*, vol.8, no.3, pp.489-508, 2017.

[4] P. Tarau and E. Figa, Knowledge-based conversational agents and virtual storytelling, *Proc. of 2004 ACM Symposium on Applied Computing*, pp.39-44, 2004.

[5] R. G. Athreya, A.-C. N. Ngomo and R. Usbeck, Enhancing community interactions with data-driven chatbots – The DBpedia chatbot, *The Web Conference 2018*, Lyon, France, pp.143-146, 2018.

[6] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998.

[7] F. Mahdisoltani, J. Biega and F. M. Suchanek, YAGO3: A knowledge base from multilingual wikipedias, *Conference on Innovative Data Systems Research (CIDR)*, 2015.

[8] W. Wu, H. Li, H. Wang and K. Q. Zhu, Probase: A probabilistic taxonomy for text understanding, *Proc. of the 2012 ACM SIGMOD International Conference on Management of Data*, pp.481-492, 2012.

[9] R. Speer, J. Chin and C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, *The 31st AAAI Conference on Artificial Intelligence*, 2017.

[10] S. Yoo, J. Song and O. Jeong, Social media contents based sentiment analysis and prediction system, *Expert Systems with Applications*, vol.105, pp.102-111, 2018.

[11] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *arXiv Preprint*, arXiv:1810.04805, 2018.

[12] Y. Zhang, V. Zhong, D. Chen, G. Angeli and C. D. Manning, Position-aware attention and supervised data improve slot filling, *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp.35-45, 2017.

[13] S. Alsubaiee, Y. Altowim, H. Altwaijry, A. Behm, V. Borkar, Y. Bu and E. Gabrielova, AsterixDB: A scalable, open source BDMS, *Proc. of the VLDB Endowment*, vol.7, no.14, pp.1905-1916, 2014.