ICIC Express Letters
Part B: Applications
Volume **11**, Number **6**, June **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 539–544

PERSON-RELATION EXTRACTION USING BERT BASED KNOWLEDGE GRAPH

Sung Min Yang, So Yeop Yoo, Yeon Sun Ahn and Ok Ran Jeong*

Department of Software
Gachon University
1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do 13120, Korea
{ ysm0622; bbusso; yeonsun0517 }@gc.gachon.ac.kr; *Corresponding author: orjeong@gachon.ac.kr

Received December 2019; accepted March 2020

Abstract. *Artificial intelligence technology has been actively researched in the areas of image processing and natural language processing. Recently, with the release of Google's language model BERT, the importance of artificial intelligence models has attracted attention in the field of natural language processing. In this paper, we propose a knowledge graph to build a model that can extract people in a document using BERT, and to grasp the relationship between people based on the model. In addition, to verify the applicability of person extraction techniques using BERT based knowledge graphs, we conduct a performance comparison experiment with other person extraction models and apply our proposed method to the case study.*

Keywords: Relation extraction, Knowledge graph, Named entity recognition

1. **Introduction.** Artificial intelligence research is developing at a rapid pace with the evolution of various hardwares that can handle data and complex operations. Artificial intelligence technology has been rapidly applied and developed especially in image processing and image recognition fields. In the field of image processing, it took a relatively short time to achieve high accuracy after artificial intelligence technology was actively used. However, the application of artificial intelligence has been studied more slowly in the human language, that is, in the field of natural language processing, and the performance enhancement has not reached the expectation of explosive performance improvement in image processing fields. The image was quickly applicable because it could be expressed in computational numbers, but natural language was considered to be difficult to training because the language was expressed in various forms according to the language type [1-3].

In recent years, various artificial intelligence technologies have been studied for natural language processing as a study of Word2vec, which expresses inputted sentences in the form of vectors, and rapid development and deep learning models in various fields related to natural language processing have achieved high accuracy.

In this paper, we apply the BERT [4] language model of Google, which shows high accuracy, to extracting person in documents. In a document such as a novel or news, a lot of words appear in the document and the meaning of each word has to be trained by the machine. At this time, the person included in the document can be extracted by applying the BERT language model.

In addition, a knowledge graph is constructed to determine the relationship between the extracted people. People have various relationships with each other, and this relationship includes various meanings such as degree and kind of relation. To find how the extracted people in the document are connected to each other, we use a knowledge graph that can connect entities and graph them.

DOI: 10.24507/icicelb.11.06.539

Relationship extraction and the knowledge graph are applied to books and publications to help the machine better understand the content itself. This is important because it can be better utilized in a content-based filtering recommendation system.

The extraction of person-relationships using BERT and knowledge graphs can extract people from various documents, especially novels, and link the relationships among people. Based on this, it is possible to use it in various ways such as analysis of relationship flow among persons, emotional analysis. In order to construct a knowledge graph based on BERT, and to verify the applicability of the person extraction technique using this, we apply the proposed method in Harry Potter I in the performance comparison experiment with other person extraction models.

2. Related Work.

The field of natural language processing is a very important field in artificial intelligence because it enables the machine to understand human language and can analyze human intention and emotion based on it. Previous studies used dictionary based processing or statistical methods to process natural language. In recent years, artificial intelligence technology has been actively studied, and it has shown high performance in the field of natural language processing based on neural network technology [1-3,5].

BERT [4] is a language model released by Google in 2018. The BERT language model was the highest performance on dataset leaderboards in 11 areas of natural language processing, including SQuAD, immediately after its release. BERT is a technique to preliminarily train general language models through unsupervised training from large corpus such as Wikipedia and to fine-tune parameters by conducting training by using datasets of specific fields such as question answering and named entity recognition, (semi-supervised learning) [5].

Among the various fields of natural language processing, named entity recognition is a field that extracts entity names such as a person, organization, location, date, and time in a sentence [6,7]. There are various kinds of entity names according to the dataset criteria that classify entity names. In this paper, we apply the named entity recognition technique to <person> among several entity names and use it to extract the person in the document.

There are various studies to extract the relation between objects. In particular, knowledge graphs represent relationships between objects based on the knowledge base such as Wikipedia. Since relationships can be expressed in a graph form, the relationship of objects can be quickly extracted even in large-scale data [8-10].

Pingle et al.'s research [11] proposed a system called RelExt that improves the Cybersecurity Knowledge Graph (CKG) using a deep learning approach in the field of relation extraction. It predicts the relationships between entity-pairs with neural networks, proving that this approach can be applied to cybersecurity as well.

There are a lot of studies to grasp the relationships among objects using knowledge graphs, but there is a limit to existing knowledge graphs in order to grasp the relationships among people in documents. Various studies use knowledge graphs to extract and analyze relationships among people, but mainly analyze the degree of intimacy between people, the degree of relationship, and the distance within a document.

In this paper, we propose a method to extract the relation by analyzing the sentence structure in which the person appears and to build and extend the knowledge graph based on this.

3. BERT Based Knowledge Graph.

We propose a knowledge graph that can extract person relations based on the BERT model which has been recently released and studied in various natural language processing fields.

Figure 1 shows the sequence of the proposed BERT-based knowledge graph. We build a training model based on BERT for extracting people and use it to extract people

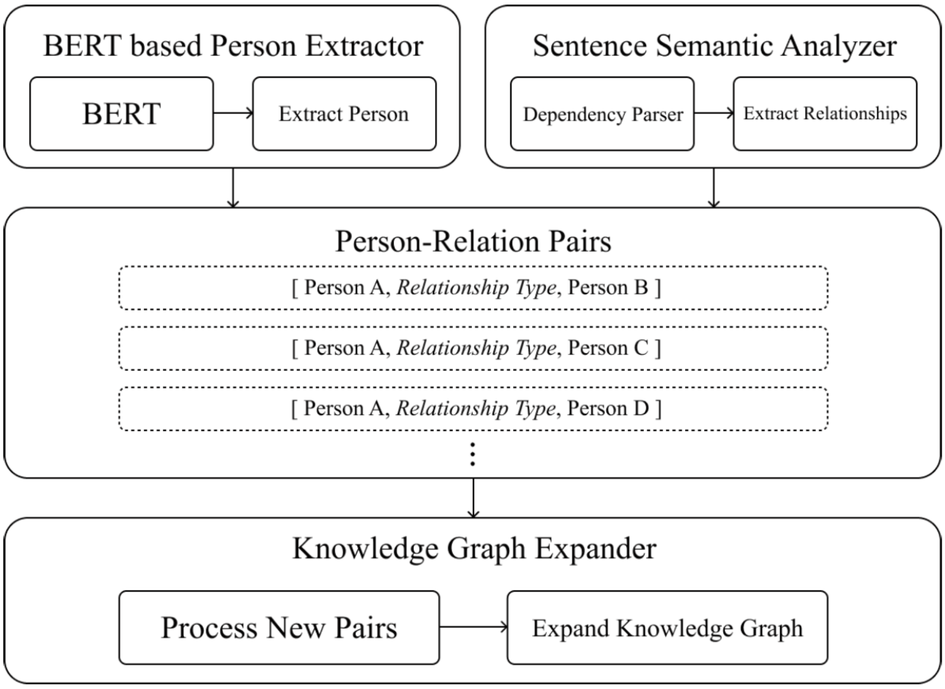
BERT based Person Extractor
BERT
Extract Person
Sentence Semantic Analyzer
Dependency Parser
Extract Relationships
Person-Relation Pairs
[Person A, *Relationship Type*, Person B]
[Person A, *Relationship Type*, Person C]
[Person A, *Relationship Type*, Person D]
Knowledge Graph Expander
Process New Pairs
Expand Knowledge Graph

Figure 1. BERT-based knowledge graph structure

in the document to create person list data. The Sentence Semantic Analyzer uses the Stanford Dependency Parser [14] to extract words and their relationships. A new person relationship pair is then extracted and added to the knowledge graph based on the person extracted by Person Extractor. This is called a Person-Relation Pair. If the new person-relationship pair is not in the existing knowledge graph, the graph is expanded while adding new ones.

3.1. **BERT based person extraction.** In this paper, we trained two models among 7 pre-trained models formally provided by BERT official. One is "BERT-base, Uncased", which has 12 layers, 768 hidden units, 12 heads, 110,000 parameters, and non-case sensitive. And the other is "BERT-large, Cased", which has 24 layers, 1,024 hidden units, 16 heads, 340,000 parameters, and a case-sensitive model.

Since the BERT model is a general purpose language model to be applied to various natural language processing fields, a pre-processing step and a parameter addition training step for named entity recognition are required. In the preprocessing step, a vocabulary for word embedding has to be generated and a tokenization process is performed using a dataset. BERT officially encourages the use of the SentencePiece [12] library to create a vocabulary. We used the tokenizer recommended by BERT for tokenization.

In the additional fine-tuning phase, the model is trained using the CoNLL 2003 [13] dataset, which is used most often in the field of named entity recognition. The four entity names that are labeled in the CoNLL dataset are Person, Organization, Location, and Misc.

3.2. **Relational extraction and generation of person-relation knowledge graph.** Knowledge graphs are knowledge-based graphs that graph-shape represents relationships between objects. It connects words and objects that can be found in a lot of data on the web and extracts the relationship and helps machine recognize this like human knowledge. Based on this, various analyses can be made possible to obtain additional meaningful knowledge.

Applying the characteristics of this knowledge graph to the document, we extract the relationships among the person extracted by the BERT language model. Using the extracted list of the person, it is filtered by the sentences in which a person was appearing, and again the sentences in which the specific person and the other person appear are commonly filtered. The structure of the extracted sentences is analyzed by using the Stanford Dependency Parser, and the relation between the two persons is extracted and newly saved. In the case of a new relationship, it is newly added to an existing graph, and in the case of an existing relationship, the relationship is weighted by adding the weight of the relationship.

4. **Case Study.** In this paper, we propose a method of constructing a knowledge graph using BERT model to extract person relation in documents. In order to extract all the people in the document, we have developed a human extraction model by training the BERT [4] model. Based on this model, the relation between the people is extracted from the sentence and expressed as a relationship graph.

Figure 2 shows an example of applying the proposed BERT-based knowledge graph to Harry Potter I. In order to apply the BERT-based person extraction model to Harry Potter I, the whole dataset of Harry Potter I is torqued with a BERT tokenizer SentencePiece library, and then create a word dictionary. The table at the top of the figure is the result of the person extraction. Using the BERT-based person extraction model in the original text, words tagged with [B-PER] are extracted as words representing the person, and a list of people is created. Actual results may include tags such as [SEP], which are used to distinguish the basic sentences of the BERT, but only the results that correspond to the person tags are displayed to show intuitively extracted results.

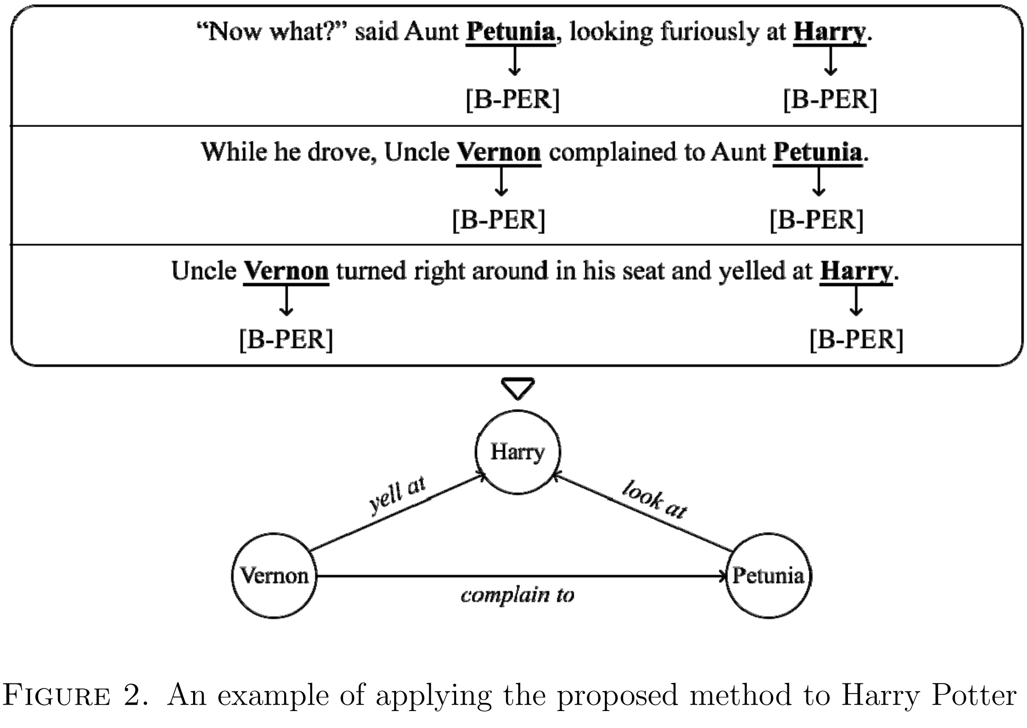
"Now what?" said Aunt Petunia, looking furiously at Harry.
[B-PER]
[B-PER]
While he drove, Uncle Vernon complained to Aunt Petunia.
[B-PER]
[B-PER]
Uncle Vernon turned right around in his seat and yelled at Harry.
[B-PER]
[B-PER]
Harry
yell at
look at
Vernon
complain to
Petunia

FIGURE 2. An example of applying the proposed method to Harry Potter I

A knowledge graph is constructed through relationship extraction for the person in the extracted person list. As shown in the lower part of Figure 2, we analyze sentences that have various people together to extract relations and display them in graph-shape.

In order to verify the performance of the most important part of a person extraction model in the proposed method, the accuracy using the F1-score was measured. We use two models "BERT-base, Uncased", "BERT-large, Cased" proposed in this paper for accurate measurement and CoNLL 2003 [13] for the training dataset. We compared the performance of the FIJZ [15] model and the bidirectional LSTM-CRF [16] model using the CoNLL 2003 testset for comparison with our model.

FIJZ [15] proposes a classifier framework that combines four different classifiers. It was the model with the highest accuracy in the CoNLL named entity recognition task. Bidirectional LSTM-CRF[16] combines Bi-LSTM (Bidirectional Long Short-Term Memory) and CRF (Conditional Random Fields). This model, proposed in 2015, greatly improves accuracy by applying a neural network to named entity recognition.

Table 1. Accuracy of person extraction model

Model	F1-score
FIJZ [15]	0.8876
Bi-LSTM-CRF [16]	0.9010
BERT-base, Uncased (ours)	0.9093
BERT-large, Cased (ours)	**0.9464**

Table 1 shows the results of comparing and testing the accuracy of person extraction models. As a result of calculating the accuracy using the same CoNLL 2003 dataset, Table 1 shows that the BERT-based figure extraction model proposed in this paper shows good results. It is common to start with an uppercase letter when writing a person name in the text. Note that the Cased model with the case-sensitive feature is much more accurate than the case-insensitive Uncased model. The F1-score of the classification model trained with BERT was measured by the following formula:

$$\mathrm{F1} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

We also applied the proposed method in Harry Potter I to checking the possibility of the proposed BERT-based knowledge graph. Because there is no person-relationship dataset in Harry Potter, it is difficult to measure accuracy. So we compared it with Harry Potter's character list on Wikipedia. A total of 89 people were found in Harry Potter I, and a total of 70 people were extracted when the people were extracted using the proposed method. Because Harry Potter is a fantasy novel, fantasy words such as talking hats, moving frames, and dragons were not recognized as a person. It would be possible to supplement the model by training the model through a lot of training data in order to regard the words that are usually perceived as objects, such as fantasy novels, as a person.

5. **Conclusions.** In this paper, we propose a knowledge graph based on BERT and propose a person extraction technique using it. We built a person extraction model using BERT, which is a pre-trained model that shows high accuracy in the field of natural language processing and extracts the relationship between people by using sentence structure analysis and knowledge graph. We verified the applicability by applying the proposed method to the novel.

Applying the proposed method to more diverse documents such as novels and news, we expect to be able to utilize relationship analysis and emotional analysis over time through the extraction of person-relations. Also, it is expected that if we use the deep learning based relationship extraction model instead of the Stanford Dependency Parser for relationship extraction, we can build a person extraction model with higher accuracy.

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 MCST (Ministry of Culture, Sports and Tourism) and KCC (Korea Copyright Commission) (Nos. NRF 2019R1A2C1008412, 2015-0-00932, 2019-CONTEXT-9500).

REFERENCES

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education Limited, Malaysia, 2016.

[2] E. Cambria and B. White, Jumping NLP curves: A review of natural language processing research, *IEEE Computational Intelligence Magazine*, vol.9, no.2, pp.48-57, 2014.

[3] A. Barr and E. A. Feigenbaum, *The Handbook of Artificial Intelligence*, Butterworth-Heinemann, 2014.

[4] 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.

[5] J. Gao, M. Galley and L. Li, Neural approaches to conversational AI, *Foundations and Trends® in Information Retrieval*, vol.13, nos.2-3, pp.127-298, 2019.

[6] D. Nadeau and S. Sekine, A survey of named entity recognition and classification, *Lingvisticae Investigationes*, vol.30, no.1, pp.3-26, 2007.

[7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, Neural architectures for named entity recognition, *arXiv preprint*, arXiv:1603.01360, 2016.

[8] 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.

[9] 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.

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

[11] A. Pingle, A. Piplai, S. Mittal and A. Joshi, RelExt: Relation extraction using deep learning approaches for cybersecurity knowledge graph improvement, *arXiv preprint*, arXiv:1905.02497, 2019.

[12] T. Kudo and J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, *arXiv preprint*, arXiv:1808.06226, 2018.

[13] E. F. Sang and F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, *arXiv preprint*, cs/0306050, 2003.

[14] D. Chen and C. D. Manning, A fast and accurate dependency parser using neural networks, *Proc. of EMNLP*, 2014.

[15] R. Florian, A. Ittycheriah, H. Jing and T. Zhang, Named entity recognition through classifier combination, *Proc. of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, pp.168-171, 2003.

[16] Z. Huang, W. Xu and K. Yu, Bidirectional LSTM-CRF models for sequence tagging, *arXiv preprint*, arXiv:1508.01991, 2015.