DR based Sentence & SPO Tuple Pair Generation for Open Information Extraction

Joonyoung Jung, Dong-oh Kang
Visual intelligence Research Section
Electronics and Telecommunications Research Institute
Deajeon, Korea
jyjung21@etri.re.kr

***Abstract*— This paper presents a method for generating sentence and SPO (Subject-Predicate-Object) tuple pairs based on dependency relations (DR) to train a DNN model for open information extraction. The proposed system extracts sentences from the OLLIE dataset that contain seed tuples in NL-based sentences with SPO tuples, performs dependency parsing on the NL-based sentences, and generates DR-based sentence and SPO tuple pairs. Consequently, approximately 230,000 DR-based sentence and SPO tuple pairs were generated from the 3 million data. To demonstrate the effectiveness of the generated DR-based learning data in natural language processing, experiments were conducted using the BERT model to recognize SPO tuples. The results showed that the performance of SPO tuple extraction was better when using DR-based learning data compared to NL-based learning data. Specifically, the average accuracy for top-1, top-3, and top-5 was 0.08, 0.25, and 0.22 higher, respectively, when using DR-based learning data compared to NL-based learning data.**

Keywords*— *dependency relation, SPO tuple, open information extraction.

I. INTRODUCTION

Research is being conducted in the field of natural language processing (NLP) to extract information from a vast amount of structured and unstructured data. The information extracted through NLP can be utilized in various applications, including expanding the large-scale knowledge bases. Therefore, it is crucial to extract knowledge efficiently and cost-effectively from a vast amount of natural language (NL) data. To extract knowledge from a large amount of NL data, it is essential to have the ability to extract knowledge not only from structured data, such as tables, but also from unstructured data, such as NL text. Extracting knowledge from unstructured data can be challenging. However, since a multitude of new and diverse information exists in the form of unstructured data, research on extracting knowledge from unstructured data is actively being studied [1]. Recently, research has been conducted on relation extraction using supervised learning approaches with neural networks. Stanovsky et al. proposed a supervised learning approach using bi-LSTM and Semantic Role Labeling models to extract tuples [2]. Shi et al. suggested the use of BERT and LSTM models for relation extraction and semantic role labeling tasks [3]. As the amount of unstructured data increases and the complexity of the model grows, the performance of SPO (Subject-Predicate-Object) tuple extraction from unstructured data can be improved. Therefore, recent NLP research focuses on enhancing the performance of SPO extraction by training complex deep learning models like BERT on large-scale learning data. However, training complex deep learning models with abundant data incurs significant costs. Therefore, in order to apply NLP to various applications, there is a need for NLP research that explores effective approaches for SPO extraction with good performance, even with smaller amounts of training data, using less complex models. To address this, utilizing the syntactic information of unstructured data instead of semantic information for SPO tuple extraction can yield good performance with smaller amounts of training data and smaller model parameter sizes. Moreover, since it focuses on learning relationships based on the structure rather than the semantics of unstructured data, the learned results can be applied to different domains [4]. To train a DNN model for SPO tuple extraction using the aforementioned advantages of syntactic information, it is necessary to generate learning data and train the DNN model. This paper proposes a method for generating learning data consisting of dependency relation (DR) based sentences and SPO tuple pairs.

The rest of the paper is organized as follows. Section II describes DR based sentence and SPO tuple pair system, while Section III describes the DR based sentence and SPO tuple pair generation. Some concluding remarks are finally given in Section IV.

II. DR-BASED SENTENCE AND SPO TUPLE PAIRS SYSTEM

The DR based sentence and SPO tuple pair data in this paper were created based on the Open Language Learning for Information Extraction (OLLIE) dataset. In OLLIE [5], using seed tuples from ReVerb [6], a bootstrapper was employed to crawl internet data and collect about 3 million pairs of NL-based sentences and SPO tuples as learning data. An example of OLLIE's training data is provided in Table I.

TABLE I. EXAMPLES OF OLLIE LEARNING DATA

	Seed Tuple	*NL Sentence*
1	god-creat-eve	Before **God created Eve**, did something go terribly wrong with Adam ?
		God created Adam and **Eve** in His image to live in fellowship with Him.
2	paypal-accept-credit card	We **accept PayPal** , ELayaway and **Credit Cards**
		This is purely because not all the web hosting providers **accept PayPal** as a payment option , as visa electron is not the most common **credit card** in theUK.
3	credit card-accept-amex	We **accept Amex** , Visa and Mastercard **credit card** payments.
		If you wish to pay by **Credit Card** we **accept** Mastercard , Visa , **Amex** and Diners Club.

In the case of a seed tuple "god-create-eve," the crawled NL sentence examples that exhibit the SPO relationship are "Before God created Eve, did something go terribly wrong with Adam?" and "God created Adam and Eve in His image

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

to live in fellowship with Him." However, in the case of a seed tuple "paypal-accept-credit card," the crawled NL sentence examples such as "We accept PayPal, ELayaway and Credit Cards" and "This is purely because not all the web hosting providers accept PayPal as a payment option, as visa electron is not the most common credit card in the UK" do not exhibit the SPO relationship. Similarly, in the case of a seed tuple "credit card-accept-amex," the crawled NL sentence examples such as "We accept Amex, Visa and Mastercard credit card payments." and "If you wish to pay by Credit Card we accept Mastercard, Visa, Amex and Diners Club." do not exhibit the SPO relationship. As observed in the examples, since there are a significant number of NL sentences in the OLLIE data where the seed tuples do not exhibit the SPO relationship, it is necessary to extract only those NL sentences where the seed tuples are in an SPO relationship for training NLP models. Additionally, to perform DR-based learning, dependency parsing (DP) needs to be performed on the NL sentences, and DR-based sentence and SPO tuple pair should be added to the DR-based learning data. The system for generating DR-based sentence and SPO tuple pairs is shown in Fig. 1.

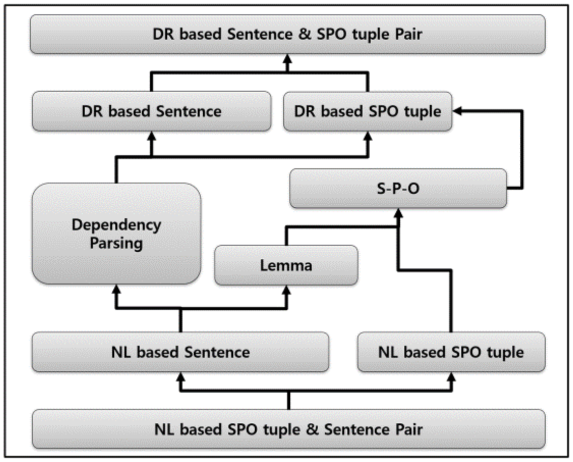


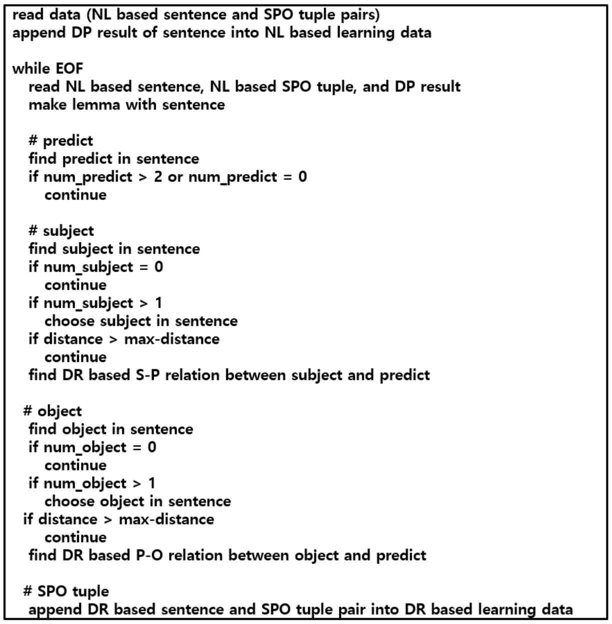
Fig. 1. DR based sentence & SPO tuple pair system.

The system for generating DR-based sentence and SPO tuple pairs can be summarized as follows:

First, extract NL-based sentence from NL-based SPO tuple and sentence pairs data. Second, perform dependency parsing on the NL-based sentences to generate DR-based sentences. Third, generate lemmas for each word in the NL-based sentences. Fourth, compare the NL-based SPO tuples with the NL-based sentences to determine which word in the NL-based sentence corresponds to each word in the NL-based SPO tuple. Fifth, utilize the results of dependency parsing and the S-P-O tuple to generate DR-based SPO tuples. Sixth, store the generated DR-based sentence and SPO tuple pairs for DNN model training.

III. DR-BASED SENTENCE AND SPO TUPLE PAIRS GENERATION

This paper extracted approximately 230,000 sentences out of around 3 million sentences, where the seed tuple exhibited the SPO relationship within NL-based sentences. The pseudo-code for extracting DR-based sentence and SPO tuple pairs is provided in Fig. 2.



```
read data (NL based sentence and SPO tuple pairs)
append DP result of sentence into NL based learning data

while EOF
  read NL based sentence, NL based SPO tuple, and DP result
  make lemma with sentence

  # predict
  find predict in sentence
  if num_predict > 2 or num_predict = 0
    continue

  # subject
  find subject in sentence
  if num_subject = 0
    continue
  if num_subject > 1
    choose subject in sentence
  if distance > max-distance
    continue
  find DR based S-P relation between subject and predict

 # object
  find object in sentence
  if num_object = 0
    continue
  if num_object > 1
    choose object in sentence
 if distance > max-distance
    continue
  find DR based P-O relation between object and predict

 # SPO tuple
  append DR based sentence and SPO tuple pair into DR based learning data
```

Fig. 2. Pseudo-code of DR-based sentence & SPO tuple pair generatoin.

The pseudo-code reads approximately 3 million sentence and SPO tuple pairs. For each sentence, it performs DP and appends the DP result to the NL-based learning data. It repeats the process of generating DR-based learning data until the last NL-based learning data is read. The generation of DR-based learning data starts by reading NL-based SPO tuple, NL-based sentence, and DP result. To handle various forms of "predict," lemma is created for each sentence. Using lemma, we search for the presence of "predict" in the NL-based sentence. If there is no "predict" or if there are more than two instances, the generation of DR-based learning data for that sentence is stopped, and the next pair of NL-based learning data, consisting of NL-based SPO tuple, NL-based sentence, and DP result, is read. If there is only one matching "predict" in the sentence, we use lemma to find the subject in the sentence that matches the subject in NL-based SPO tuple. If there is no matching subject, we finish generating DR-based learning data for that sentence and move on to the next pair of NL-based learning data. If there are multiple matching subjects, we select the subject with the least distance to the predict in both the NL-based sentence and DP result. If the distance between the selected subject and predict is greater than the maximum distance, we finish generating DR-based learning data for that sentence and move on to the next pair of NL-based learning data. If the distance between the selected subject and predict is less than the maximum distance, the DR-based S-P relation between the selected subject and predict is chosen. For the DR-based P-O relation, it is found using the same method as the DR-based S-P relation. Once the DR-based S-P relation and P-O relation are found, the DR-based sentence and SPO tuple is added to the DR based learning data.

In the case of the seed tuple "god-create-eve" and the NL-based sentence "God created Adam and Eve in His image to live in fellowship with Him.", an example of the generated DR-based sentence and SPO tuple pair is as follows:

The result of performing DP on the NL-based sentence "God created Adam and Eve in His image to live in fellowship with Him." is shown in Fig. 3.

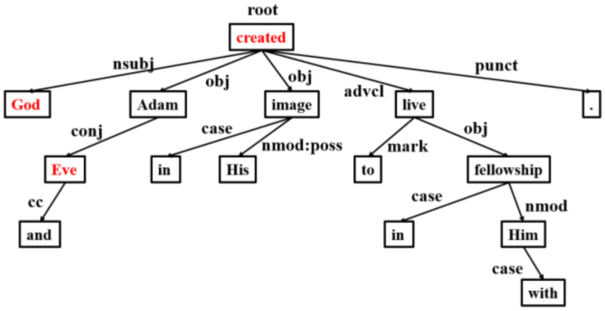


Fig. 3. DP result of "God created Adam and Eve in His image to live in fellowship with Him.".

After performing DP, the generated DR-based sentence is "nsubj root obj cc conj case nmod:poss mark case object case nmod punct", and the DR-based SPO tuple is "nsubj-predicat-(obj-conj)". This DR-based sentence and SPO tuple pair is stored together as part of the DR-based learning data.

To demonstrate the effectiveness of the generated DR-based learning data in NLP, experiments were conducted using the BERT model to recognize SPO tuples. The results obtained from applying the BERT model to the NL-based sentence and SPO tuple pair data are shown in Fig. 4. The average accuracy for top-1, top-3, and top-5 is 0.55, 0.68, and 0.75, respectively. The accuracy of SPO tuple extraction exhibits variability across different tests.

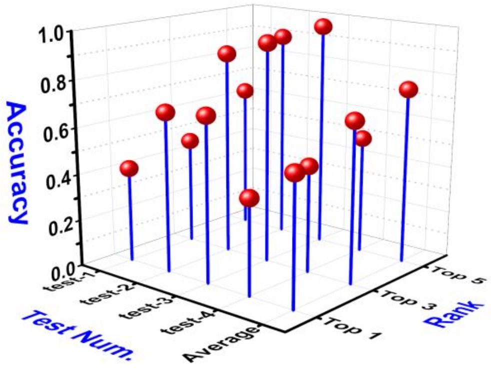


Fig. 4. NL-based SPO tuple recognition with BERT.

The results obtained from applying the BERT model to the DR-based sentence and SPO tuple pair data are shown in Fig. 5.

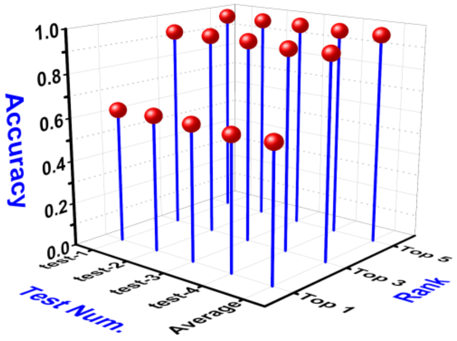


Fig. 5. DR-based SPO tuple recognition with BERT.

The average accuracy for top-1, top-3, and top-5 is 0.63, 0.93, and 0.97, respectively. The accuracy of SPO tuple extraction is high at top-5, reaching 0.97, and shows stable performance with little variation in accuracy across different tests. According to the experimental results, it can be observed that using DR-based learning data for SPO tuple extraction yields more stable and better performance compared to using NL-based learning data. Furthermore, training with DR-based sentence and SPO tuple pair data, which utilizes the structural relationships within NL-based sentences, allows for applying the learned DNN model from one domain to another without the need for separate training for each domain. This makes it valuable for learning data in the field of information extraction in NLP, as it can be applied effectively across different domains.

IV. Conclusions

The paper proposes a method for generating DR-based sentence and SPO tuple pairs to train a DNN model for SPO tuple extraction using syntactic information. The DR-based sentence and SPO tuple pair system extracts NL-based sentences from the OLLIE dataset that contain seed tuples in NL-based sentences with SPO relations, performs dependency parsing on NL-based sentences, and generates DR-based sentence and SPO tuple pairs. Consequently, approximately 230,000 DR-based sentence and SPO tuple pairs were generated from the 3 million data. To demonstrate the effectiveness of the generated DR-based learning data in NLP, experiments were conducted using the BERT model to recognize SPO tuples. The results showed that the performance of SPO tuple extraction was better when using DR-based learning data compared to NL-based learning data. Specifically, the average accuracy for top-1, top-3, and top-5 was 0.08, 0.25, and 0.22 higher, respectively, when using DR-based learning data compared to NL-based learning data.

Acknowledgment

This work was supported by Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government. [23ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System]

References

[1] Z. Jiang, P. Yin, and G. Neubig, "Improving open information extraction via iterative rank-aware learning," *Association for Computational Linguistics*, Florence, Italy, pp. 5295–5300, Jul. 2019.

[2] G. Stanovsky, J. Michael, L. Zettlemoyer, and I. Dagan, "Supervised open information extraction," *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, USA, pp. 885–895, Jun. 2018.

[3] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," arXiv: 1904.05255, Apr. 2019.

[4] J. Jung, "DG-based SPO tuple recognition using self-attention M-Bi-LSTM," *ETRI Journal*, vol. 44, issue 3, pp. 438-449, 2022.

[5] M. Schmitz et al., "Open language learning for information extraction," *Empirical Methods in Natural Language Processing*, Jeju, Rep. of Korea, pp. 523–534, Jul. 2012.

[6] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," *Empirical Methods in Natural Language Processing*, Edinburgh, UK, pp. 1535–1545, Jul. 2011.,