SRDF: Korean Open Information Extraction using Singleton Property

Sangha Nam, Younggyun Hahm, Sejin Nam, and Key-Sun Choi

Semantic Web Research Center, KAIST, Korea
{nam.sangha, hahmyg, namsejin, kschoi}@kaist.ac.kr

Abstract. In this paper, we propose a new Korean Open Information Extraction system so-called SRDF. The SRDF system has been designed to effectively extract reified triples from Korean natural language texts based on the use of singleton property and other natural language processing techniques such as part-of-speech tagging and chunking. The SRDF system is the Open Information Extraction system that enables extracting a multiple number of triples from a single sentence via reification.

1 Introduction

Traditional Information Extraction (IE) thus far has been relying heavily on human intervention of hand-crafted rules and hand-tagged training data. In recent years, on the other hand, Open IE based on self-supervised learning has become more strongly suggested to overcome such a limitation, and it is now possible to process massive text corpora without having to require much human effort. TextRunner [1], WOE [2] and ReVerb [3] are some of the most representative examples of Open IE systems that offer excellent performance in automatically extracting structured information from unstructured natural language texts. Unfortunately, however, these systems cannot guarantee the same level of performance on languages other than English. For that reason, the Chinese Open IE system as an instance is currently being actively researched [4]. In addition to this, these systems also fall short of representing multiple relationships between argument(s) and relation(s) within a sentence, since they are designed to focus primarily, or rather restrictedly, on binary extractions. In other words, the recent Open IE systems can extract only one triple with a single argument and relation respectively per sentence, whereas many of the statements, especially those describing an event, are generally inclusive of more than one argument such as time and location, and/or two or more relations. This indeed has been one of the most principal challenges remained to be addressed in the study of Open IE.

Throughout the following sections of this paper, we introduce SRDF, the new Korean Open IE system, in much greater details. Since the Korean language, in a variety of respects, has uniquely different grammatical structures and the system of postposition and word spacing compared to other languages like English and Chinese in particular, our team has been devoted to develop a new Open IE system specially designed to meet the characteristics of Korean. We, at the same time, have also strived to build a system through which multiple relationships between argument(s) and relation(s) within a sentence can be extracted by using singleton property – the new method of reification [5]. Taking the singleton property approach to extracting reified triples from Korean natural language texts is to minimize the number of triples, and to further allow the results of our system to be compatible with well-known knowledge bases such as DBpedia and YAGO.

2 Korean Open Information Extraction using SRDF

SRDF simply receives as input a Korean text corpus and returns an extracted set of triples expressed in the form of singleton property. The system of SRDF operates through three steps of procedure in total that are "preprocessing", "argument and relation detection", and "triple generation" as described below.

2.1 Preprocessing & Argument and Relation Detection

When a Korean sentence is given as input, the SRDF system performs part-of-speech (POS) tagging and chunking first, as preprocessing.

The POS-tagged and chunked Korean sentence is then passed on to the next stage of argument and relation detection. This stage literally is to detect argument(s) and relation(s) from the given Korean sentence, and is further divided into three smaller steps similar to other Open IE systems based on self-supervised learning.

- *Labeling*: At this stage, the preprocessed Korean sentence gets automatically labeled based on three important factors that are "the POS tag patterns", "the position of words in sentence", and "the postposition(s) within the sentence".
- *Learning*: An argument detection model and a relation detection model are learned here using decision tree. The former model uses "lemma", "POS tag", "length of the sentence", "start position of argument", "end position of argument", "next lemma" and "next POS tag" as features, and the features that the latter uses include "lemma", "POS tag" and "postposition".
- *Extracting*: Once a Korean sentence is received as input, the relation detection model in the SRDF system classifies whether a certain word in the given sentence is a relation or not, while the argument detection model classifies whether the word is a subject or an object. After that, they return all the classification results that are necessary for the next step of triple generation, including "the postposition(s) of detected argument(s)" and "the position of detected argument(s) and relation(s) in sentence".

2.2 Triple Generation

Our team has studied not only "how to extract information from Korean sentences" but "how to generate triples for representing multiple relationships between argument(s) and relation(s) within a sentence" as well. For example, the sentence ``Barack Obama was awarded the Nobel Prize in 2009.'' is including multiple relationships shared between one relation and two arguments, thus it is ideal that two triples should be generated like <Barack Obama, award, the Nobel Prize>, <Barack Obama, award, in 2009>. However, when we perform information extraction on the above-mentioned sentence using the ReVerb [3] program for instance, only one triple <Barack Obama, award, the Nobel Prize> is returned excluding the fact that Barack Obama was awarded "in 2009". In order to address this problem, we have adopted an approach that grafts the concept of singleton property onto Open IE.

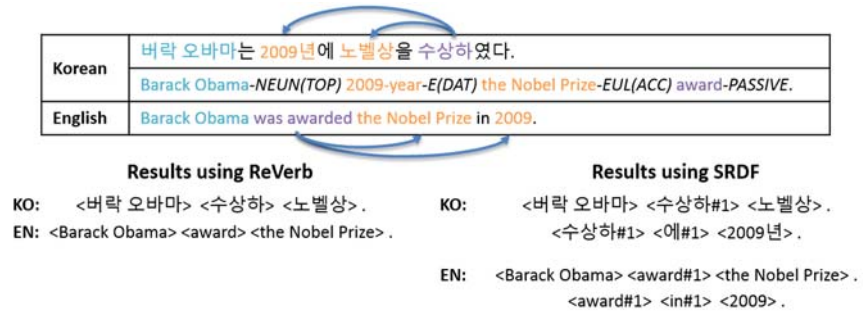

Fig. 1. Example of Korean Open Information Extraction using SRDF

As shown in Fig. 1 (blue = subject, orange = object, purple = relation), the method of generating triples using SRDF is as follows:

1. *Identify the association between argument(s) and relation(s) based on the position of words in sentence.* – Korean sentences have a different structure from English. Whereas an English sentence typically has a Subject-Relation-Object word order, the word order of Subject-Object-Relation is more common in Korean. In this light, the SRDF system can infer that the object(s) in Korean sentences are associated with the relation(s) located on the right side of them, and vice versa. In effect, the objects "`the Nobel Prize`" and "`2009-year`" are associated with the relation "award" on their right-hand side, as shown by the blue curved arrows in Fig. 1.
2. *Identify whether the postposition(s) attached to the object(s) of the sentence is accusative.* – In Korean, it is also common that almost every object is attached with a postposition, and the postposition is considered a very important factor when understanding syntax of the sentence. Among various postpositions, accusative postposition "`EUL`" specifically indicate that the relation of the sentence is a transitive verb. When the postposition attached to the object of the sentence is accusative, the SRDF system generates a triple with the following form in general `<Barack Obama, award#1, the Nobel Prize>`, in which the relation "`award`" is attached with a provenance #1. In other cases where the relation of the sentence is *not* a transitive verb and the postposition attached to the object is *not* accusative, triples are made by the SRDF system in the anonymous form of `<subject, relation#provenance, ANOMYOUMS>`. This method has an advantage of enabling representation of sentences with no object in the form of triple.
3. *Generate reified triples using remained objects.* – The main triple `<Barack Obama, award#1, the Nobel Prize>` has been made in the previous step and, at this stage, "`2009-year-E`" should be reified. When generating a reified triple, the SRDF system situates the relation of the main triple as the subject of the reified triple, places the postposition as the relation, and lets the object be the object as `<award#1, E#1, 2009-year>` for instance.

3 Experiment

The performance of SRDF system has been evaluated by application to 100 Korean sentences randomly sampled from the web as a testing data set. The evaluation results have been assessed by two human evaluators based on the two criteria of Detection – how precisely the SRDF system has detected the argument(s) and relation(s) from the given sentence – and Triple Generation – how accurately the reified-triple has been generated from the detected argument(s) and relation(s) –. The results and error statistics are presented in Table 1 below. As shown in Table 1, the SRDF system is of an excellent capability of both detecting argument(s) and relation(s) and generating triples, where the performance of triple generation is relatively 18% lower. Having thoroughly examined the failed sentences, we found out that most errors occur in the course of detection followed by POS-tagging, and the least errors are made during the process of reification.

Table 1. Performance Evaluation and Error Statistics of SRDF

Criteria	Performance			Error Statistics		
	Precision	Recall	F1-score	POS	Detection	Reification
Detection	0.81	0.86	0.83	0.15	0.74	0.11
Triple Generation	0.66	0.65	0.65			

4 Conclusion

In this paper, we have demonstrated the feasibility of extracting structured information from Korean natural language texts without any human intervention. We have also proposed a novel method of combining Open IE with the singleton property technique in representation of multiple relationships between argument(s) and relation(s) within a sentence. Our project is still ongoing in active progress, and it is with great expectation for our forthcoming researches to more technically expand the scope of our project. All the expected accomplishments of the next phases of our project work will be made publicly available through the website at http://143.248.135.216:8080/SRDFREST/index.htm.

Acknowledgement. This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. R0101-15-0054, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

References

1. Etzioni, O., et al.: Open Information Extraction from the Web. Communications of the ACM 51(12), 68-74 (2008)
2. Wu, F., and Weld, D. S.: Open Information Extraction using Wikipedia. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (2010)
3. Etzioni, O., et al.: Open Information Extraction: The Second Generation. IJCAI (2011)
4. Tseng, Y. H., et al.: Chinese Open Relation Extraction for Knowledge Acquisition. EACL (2014)
5. Nguyen, V., Olivier B., and Amit S.: Don't Like RDF Reification? Making Statements about Statements using Singleton Property. In Proceedings of the 23rd international conference on World wide web (2014)