Information Extraction Using Distant Supervision and Semantic Similarities

Youngmin PARK, Sangwoo KANG*, Jungyun SEO
Department of Computer Science and Engineering, Sogang University, 04107, Republic of Korea
**Corresponding author: gahng.sw@gmail.com*

***Abstract*—Information extraction is one of the main research tasks in natural language processing and text mining that extracts useful information from unstructured sentences. Information extraction techniques include named entity recognition, relation extraction, and co-reference resolution. Among them, relation extraction refers to a task that extracts semantic relations between entities such as personal and geographic names in documents. This is an important research area, which is used in knowledge base construction and question and answering systems. This study presents relation extraction using a distant supervision learning technique among semi-supervised learning methods, which have been spotlighted in recent years to reduce human manual work and costs required for supervised learning. That is, this study proposes a method that can improve relation extraction by improving a distant supervision learning technique by applying a clustering method to create a learning corpus and semantic analysis for relation extraction that is difficult to identify using existing distant supervision. Through comparison experiments of various semantic similarity comparison methods, similarity calculation methods that are useful to relation extraction using distant supervision are searched, and a large number of accurate relation triples can be extracted using the proposed structural advantages and semantic similarity comparison.**

***Index Terms*—relation extraction, unsupervised learning, distant supervision, information extraction, natural language processing.**

I. INTRODUCTION

In recent years, social media has been expanded and the use of smartphones is increasing, creating various types of unstructured data explosively. To pace with this trend, a large number of studies have been conducted on how to extract useful information from unstructured data.

Information extraction is one of the main research tasks in natural language processing and text mining that extract useful information from unstructured sentences. Information extraction techniques include named entity recognition, relation extraction, and co-reference resolution. Among them, relation extraction refers to a task that extracts semantic relations between entities such as personal and geographic names in documents, and it is a valuable research area that is used in knowledge base construction and Question and Answering (Q&A) systems.

Relation extraction is a task to extract semantic relations between two named entities, which refer to personal, geographic, or organizational names. Such named entities cannot be constructed as a dictionary-like type because they may be newly created or modified over time. Accordingly, studies have been conducted to extract semantic relations from newly created named entities using already known named entity information. Many studies on relation extraction have been performed on supervised learning methods through research on features and kernels, but in recent years, semi-supervised learning methods have been conducted to reduce financial and time expenditures for supervised learning. In recent years, much attention has been paid to a semi-supervised learning method using a knowledge base called distant supervision. Distant Supervision is a information extraction method that first proposed in the biomedical domain [1]. Distant supervision uses *Freebase* as a knowledge base that stores semantic relations between entities based on entities [2]. A tagging of collected unstructured sentence data is performed automatically using a knowledge base. More recently, various methods that apply a distant supervision learning technique to relation extraction have also been proposed. Multilateral approaches have been attempted to reduce the previously mentioned errors and promote diversification of the knowledge base. A direct approach to knowledge bases and error analysis on a learning corpus after a distant supervision assumption has been also attempted in various ways.

This study also approached the distant supervision assumption through semantic analysis. That is, we aimed to find improvements on the distant supervision assumptions via the semantic approach method using refined semantic analysis dictionaries such as WordNet from one-dimensional corpus-oriented analysis. In other words, we propose a structured method to improve the limitations of semantic relation dictionary definitions and to reduce errors by adding a semantic approach to the distant supervision assumption. Using various natural language processing-based techniques, sentences are analyzed and semantic meanings between entities are extracted. Then, meanings are converted into a cluster, and sentences that have the same semantic meaning are searched and found through the semantic similarity analysis process followed by matching them with those in a knowledge base, thereby having a learning structure in which sentences that have preferred semantic relations are found using a training data set.

II. RELATED WORKS

Studies of supervised machine learning-based relation extraction that employs statistical classifier with annotated training data can be classified it as a feature-based or kernel-

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2013R1A1A2010190) and this work was supported by the ICT R&D program of MSIP/IITP. [R0126-15-1112, Development of Media Application Framework based on Multi-modality which enables Personal Media Reconstruction]

Digital Object Identifier 10.4316/AECE.2016.01002

based. The feature-based relation extraction was proposed by [3-4]. This study employs features extracted from natural language understanding such as POS tagger, syntactic parsing, named-entity tagger. In the kernel-based approach, a relation extraction model uses shortest path on a dependency graph extracted by dependency parser [5], and tree kernel for relation extraction was presented by [6].

On the one hand, semi-supervised or unsupervised approaches have been actively studied for relation extraction. An open information extraction system was presented by [7], this model extracts relation using heuristic matches between *Wikipedia infobox* and corresponding sentence. Meanwhile, a relation extraction method extends relation triples using small seed data in open web environment [8].

A recent trend in relation extraction is methodology to employ untagged big data based on semi-supervised learning. Especially, the study of Distant Supervision was represented that generates training corpus for relation extraction from raw corpus with *Freebase* as knowledgebase [9]. In the experiment, it shows significant improved performance. But results of Distant Supervision method have flaw that generated corpus may include wrong labeled relations. To solve this problem various studies have progressed. The model to eliminate wrong labeled relations was proposed by [10], and those experiments show that patterns effectively detect wrong relations. And the study of Augenstein improved a precision of Distant Supervision precision through a method to discard unreliable seed data [11]. Also the method to detect valid entity types in relation was presented by [12]. Moreover, a study of multiple relation extraction method was introduced by [13]. In addition, various Distant Supervision methods have been studied in relation extraction [14-16].

Distant Supervision variously has modified to other tasks besides relation extraction. Distant Supervision was applied to extract temporally anchored relation [17] and slot filling system [18]. In named entity recognition, a study of Kim shows that Distant Supervision effectively generates corpus to type of person, location and organization [19].

III. Relation Extraction Model

This section introduces the previously discussed contents in detail as follows: First, we describe a pre-processing technique. And second, we propose an automatic training data configuration for cluster candidate extraction and semantic similarity-applied merge. At last, a detailed explanation about how to configure the binary classifier is presented.

A. System Architecture

We first present an overall structure of the proposed method, to help us understand the entire system. As Figure 1 shows, a pre-processing step is performed with sentence data when sentence data are entered as input, and then cluster candidate extraction and semantic relation merge steps are conducted on the basis of the pre-processing outcome to configure training data automatically. Using the results of the merge step, a cluster that corresponds to a class of a knowledge base is found and connected to perform tagging work.

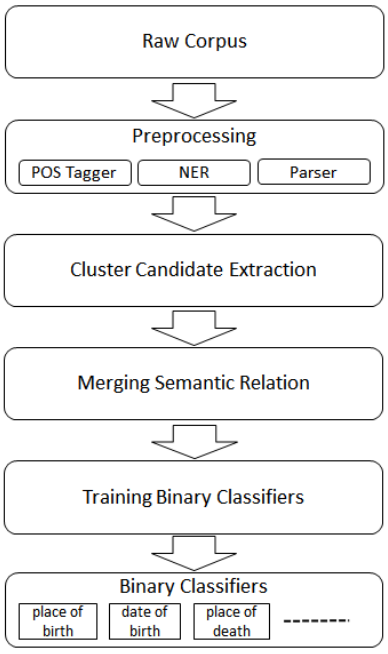

Figure 1. Procedure of the Proposed System

This tagging work is more effective in semantic accuracy than using only the distant supervision assumption introduced in the related study, thereby obtaining accurate training data, and having sentences that include more useful features than those using only the distant supervision assumption. The reason for this is that many useful corpus can be extracted from the extraction process and semantically close sentences can be collected using semantic analysis during the merge work. On the basis of the configured training data, which is considered as a correct answer, a binary classifier is created for each semantic relation. Using the created binary classifier, additional entities that are corresponded to the semantic relation can be found for newly introduced data. The reason is that the Abstract should be understandable in itself to be suitable for storage in textual information retrieval systems.

B. Preprocessing

A pre-processing task is needed to extract required information from unrefined sentences to use them in the core part of the system. It refers to a task that helps reading various pieces of information from sentences using natural language processing techniques used in the natural language processing field.

This study uses a Part-Of-Speech (POS) Parser, Dependency Parser, and Named Entity Recognizer. Using them, a sentence POS, dependency, and named entity tag are attached automatically. Tag information in a preprocessing task is used variably from later training data configuration to the classifier's features. There are five types of a named entity that can be obtained through the preprocessing task: person, location, organization, date, and degree. They are attached automatically through the named entity recognizer and pre-processing task. A morpheme and a POS can be obtained via the POS parser, whereas the dependency between morphemes can be obtained via the dependency parser.

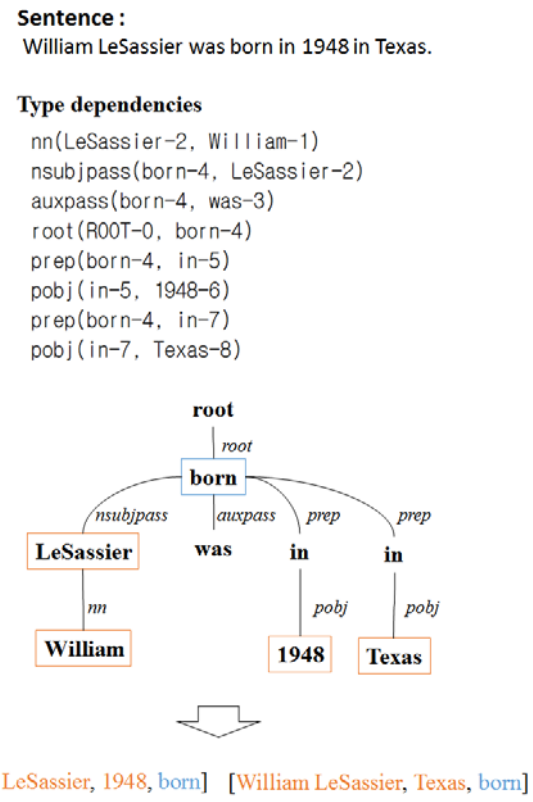

Figure 2. An example of Candidate Extraction

C. *Cluster Candidate Extraction*

In this paper, all processes that compose training data are conducted automatically. All processes are conducted without prior definition on semantic relations at the initial state. A triple is extracted in the format of <named entity 1, named entity 2, semantic relation> using pre-processing information of the sentence data. Each triple is represented as a single cluster, and clusters that have the same semantic relation are merged through a merging task in which semantic relations are applied. Through this, various representations that have the same semantic relation can be found.

The extraction of semantic relation candidates is conducted with sentences that contain more than two entities after the named entity recognizer recognizes named entities. As a semantic relation between two entities, a triple <named entity 1, named entity 2, semantic relation> is extracted by defining that a semantic verb existed in the shortest path between two entities as a semantic relation candidate through a study by Bunescu et al. [5].

As shown in Figure 2, dependency parsing results are used to extract a triple. For example, when a path between “William LeSassier” and “1948” is traced, a semantic relation "born" is extracted so that a triple is extracted in the format of <William LeSassier, 1948, born>. A sentence that contains more than three entities can extract a triple for all combinations between entities. If a sentence contains K entities, 2K triples can be extracted.

Once a triple is extracted in the initial cluster candidate extraction step, triples of <named entity 1, named entity 2, semantic relation> are converted to a cluster type. A triple is converted into a format of <entity type 1, entity type 2, semantic relation> according to a pair of entities: entity type 1 and entity type 2. By binding triples that are represented as the same semantic relation in the same entity type pairs, a single initial cluster is created. For example, a <William LeSassier, 1948, born> triple is converted to <PERSON, DATE, born>, and all sentences in which the same triples are extracted are made into a single cluster. Here, even if the semantic relations are the same with each other, they can be in different clusters, depending on their entity type. For example, as Figure 2 shows, despite that fact that two triples of the same semantic relation "born" were extracted, they can be in different clusters, depending on entity types. Two triples are converted into <PERSON, DATE, born> and <PERSON, LOCATION, born>, respectively, which belong to different clusters.

D. *Merging semantic relations*

As explained previously, extracted initial triple candidates are present as a cluster format. In the semantic relation merging step, clusters of the same semantic relation are merged. Through this task, clusters that are determined to have the same semantic relation as a result of synonym and paraphrase are merged, although they were classified initially as a different cluster.

To merge semantic relations, similarity between two clusters should be calculated. The semantic similarity calculation has been studied in a variety of ways. In this study, we compared and analyzed represented similarity comparison methods. In next section, we explain in detail the similarity comparison methods that we used.

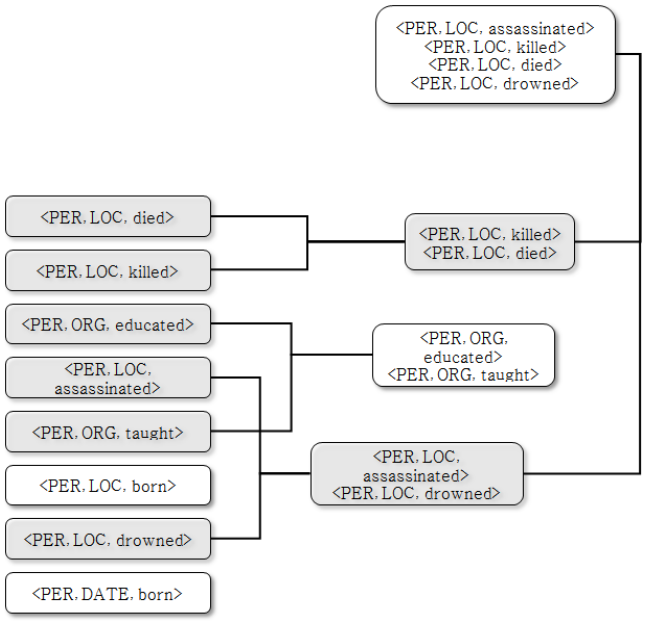

Figure 3. Merging Semantic Relation

Figure 3 shows a simple example merging step according to semantic relation similarity. Similar to the Hierarchical Agglomerative Clustering method, similarity between clusters is calculated, and two clusters that have a high similarity are found and merged in this clustering. In this study, similarity between two clusters is calculated to merge them in the semantic relation merging step. Each cluster is produced within collected learning corpus. Because frequencies of usage in sentences are different, the number of sentences included in a cluster differs. To compare the

semantic similarity between clusters, four methods were used to apply various methods such as comparison between contained sentences and semantic comparison between represented semantic verbs. Section 3.4 discusses these methods. The final outcome of the semantic relation merging step is to find synonyms, thesauruses, and paraphrases, thereby storing them as a cluster format. In the clustering step, a threshold of semantic similarity may be set as a termination condition, or the number of final clusters required can be set. In this study, we used a threshold of similarity.

Studies on similarity comparison between words or specific senses have been conducted for word sense disambiguation. In this study, we find expressions that have relations such as synonyms, thesaurus, and paraphrases between triple clusters, thereby determining whether they are produced into a single cluster that represents the same semantic relation.

A measurement of semantic similarity has been variably studied in natural language processing and information retrieval fields. The semantic meaning of words is determined, depending on various factors such as circumstances or contexts, and even the same semantic meaning can be expressed variably. In this study, we compared and analyzed several methods that can be applied to solve the previous problems among these semantic similarity measurements. By summarizing techniques that are applied to compare semantic verbs or real corpus, four methods were found, compared, and analyzed.

1) WordNet

WordNet is a lexical database of English developed by the Cognitive Science Laboratory of Princeton University. It classifies English words into sets of synonyms and records various semantic relations using combinations of synonym and antonym dictionaries as well as a vocabulary dictionary. It was started from an interest in machine translation initially in 1985 and since then has been gradually expanded.

This study uses synonym information, verb group (Synset) information, and distance similarity between two concepts, which can be acquired in WordNet. Representative similarity calculation methods in WordNet are the edge-based measurement method and information content measurement method. This study uses an edge-based measurement method. Among the edge-based measurement methods, we used a value calculated via (1).

$$distance = \frac{minDistToCP}{distFromCPToRoot + minDistToCP} \tag{1}$$

where CP is a common parent, which refers to a common ancestor concept of two concepts. *minDistToCP* is the shortest path to CP between two concepts and *distFromCPToRoot* is a distance from Root to CP in WordNet. This equation is defined to have higher similarity between two concepts if the distance between two concepts is shorter and the depth is deeper. If the outcome of this equation is lower, the similarity between the two concepts is higher.

2) Cosine similarity

Cosine similarity is a widely used method to compare similarity between documents. A cosine similarity measurement method measures how similar two vectors are with each other by measuring a cosine angle between two vectors. Each document is expressed as a vector, and cosine similarity between document vectors is measured to determine whether two documents are similar, the closer the direction of the two vectors, the closer the similarity.

$$cos\theta = \frac{d_i \bullet d_j}{\|d_i\|\|d_j\|} = \frac{\sum_{i=1}^{n} d_i \bullet d_j}{\sqrt{\sum_{i=1}^{n}(d_i)^2} \times \sqrt{\sum_{i=1}^{n}(d_j)^2}} \tag{2}$$

The cosine similarity between document d_i and d_j vectors is calculated by (2). In this study, while conducting a clustering step in consideration of semantic relations, each cluster should contain candidate sentences that correspond to the cluster. Cosine similarity between clusters is calculated using a method that is similar to the similarity comparison between documents.

3) Pointwise Mutual Information

PMI, along with the cosine similarity, is a method to calculate similarity directly within corpus. There are also many improved methods of PMI, but this section uses a basic PMI method to analyze the PMI-based similarity calculation method. In the next chapter, a similar but improved PMI method is described.

A PMI method is used to calculate the inter-relation between two words. As an equation of calculating similarity between two words in collected documents, the probability of occurrence of each word in a document and the probability of occurrence of two words at the same time are calculated. Two words will have high similarity if they are used in a sentence frequently at the same time.

$$PMI(word_1, word_2) = log_2\left(\frac{P(word_1, word_2)}{P(word_1)p(word_2)}\right) \tag{3}$$

In (3), $p(word_1)$ is the probability of occurrence of word 1 and $p(word_2)$is the probability of occurrence of word 2. $p(word_1, word_2)$ is the probability of co-occurrence of two words. The degree of inter-relation between two words is calculated using (3) and in this study, they were used as a measure to calculate similarity.

4) Collocation

Collocation is a frequent combination of a particular word with another word or words to express a specific meaning. Lin [20] proposed a method to calculate similarity between words using peripheral features of words in a study on collocation of specific words from corpus. To calculate similarity between two words, subjects, verbs, and adjectives used with specific words are designated as features, and an information amount of each feature is calculated, thereby calculating similarity between words.

In this study, a collocation similarity calculation method [20] was applied to semantic relation words of clusters, thereby using peripheral words as features to compare semantic relation similarity. That is, with this method, semantic verbs that share features of a high amount of information as a result of applying the same equation used in PMI will have high similarity. This method is a typical semantic comparison analysis method that is dependent on corpus.

E. Relation extraction with binary classifier

In general, a classification process for relation extraction is to classify a semantic relation class and a semantically unrelated class using a multi-class classifier or to classify whether semantic relations are included or not using a binary classifier and to reclassify sentences classified as having semantic relations to determine which semantic relations they belong to using a multi-class classifier.

1) Binary classifiers

In this study, semantic relations are not defined in advance, and each final cluster result is configured using a classifier. Finally, a semantic meaning is determined by finding a cluster that corresponds to the semantic relation class defined in a knowledge base. This step-by-step configuration is in contrast to existing methods in previous studies that define semantic relations in advance, cannot know how many classes are finally determined, and increases the amount of final classes because a large number of clusters may be created, depending on corpus or knowledge base.

The configuration up until here may define each of the final clusters as a single class by refining them even without a knowledge base. If this type of configuration is used, a method of how to find a correctly determined cluster should be designated. In this study, distant supervision is applied along with a knowledge base so that a cluster that has the largest number of entity pairs in a knowledge base is determined as a corresponding class.

Because the selection of such structural adoption, it is unclear how many classes are finally produced, hence a binary classifier as per class was used to classify classes rather than a multi-class classifier. Each classifier is a binary classifier that classifies whether sentences belong to the class or not.

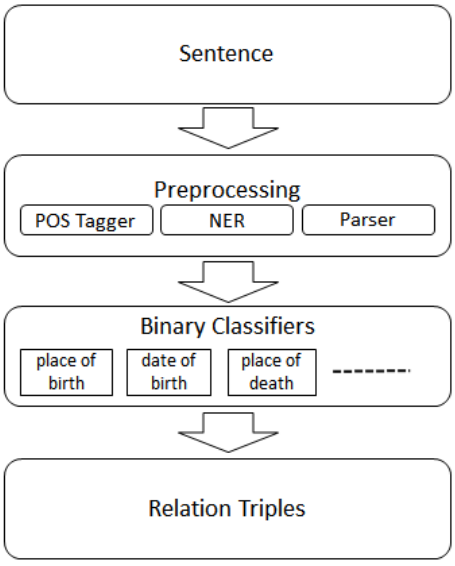

Figure 4. Relation Extraction with Binary Classifiers

Figure 4 shows the process of classification using a binary classifier. The semantic relation merging step in Figure 3 corresponds to the process where clusters are created in Figure 3, and each cluster is used as training data to train a binary classifier that classifies a corresponding cluster. Here, sentences used in other classifiers and sentences that do not include pairs of the named entities as many as the number of learning sentences are used as incorrect data for training purposes.

2) Features

In this study, three types of features are used mainly for relation extraction: named entity, lexical, and dependency. Basically, the feature is similar to the relation extraction features used in a study by Mintz et al. [9], in which relation extraction was attempted using distant supervision. A named entity feature consists of strings of two named entities; a lexical feature consists of a unigram and bigram of morpheme and POS; and a dependency feature consists of information produced from dependency parsing.

It is highly important for relation extraction to know information about named entities because it aims to identify semantic relations between two named entities. The length of the strings, types, and constituent characters would be different depending on the named entity types such as persons, locations, or date. There is a unique characteristic of each named entity type such as the inclusion of strings, as in Feb or October, person names, locations, and organization names, although exceptional cases can be found. N-gram features were used for named entities to use the previous characteristics in training data. A type of named entities in pre-processing results was removed from the training data to prevent propagation of errors in the named entity recognizer.

In the lexical feature, morphemes of constituent strings in sentences excluding named entities and their POSs were used. According to the importance of the study, results by Zhou et al. [4], front and rear strings of two named entities which were useful for relation extraction were removed, and morphemes in sentences and their POS's unigram and bigram were used as features.

As for dependency features, the features that contain two entities and strings between them were used. As with the lexical feature, dependency relations between sentences were used. Because the Stanford Dependency Parse representation method uses a format of *nsubj*(died-2, Nicolson-1) for strings, there is a risk that string morphemes may be included in features. To solve this problem, only location information, which was not a string type, was included to change the above format into a format of *nsubj*(2,1).

TABLE I. BINARY CLASSIFIER FEATURES

Feature	Extracted Features
Entity 1	John Lennon
Entity 2	1980
word uni-gram	John, Lennon, was, killed, in, ...
word bi-gram	John Lennon, Lennon was, was born, born in, ...
POS uni-gram	NNP, VBD, VBD, IN, CD, ...
POS bi-gram	NNP VBD, VBD VBN, VBN IN, IN CD, ...
Dependency	nsubjpass(4,2), auxpass(4,3), prep_in(4,6), ...

Table I shows an example of feature creation using pre-processing with respect to the following sentence "John Lennon was killed in 1980 in New York."

IV. Experiment and Results

A. Experiment environment

In this chapter, learning and experimental corpus and experimental results are presented to evaluate the proposed system. To evaluate the system, 5-fold cross validation was conducted with respect to triples in a knowledge base, and similarity techniques and a distant supervision assumption were mixed in the experiment. Two results between the existing distant supervision assumption and the proposed semantic similarity were compared and analyzed.

Experimental data used in this paper is corpus that is disclosed for relation extraction experiments by Google. This data has five semantic relation classes. These corpora consist of 100,000 sentences and 59,576 pairs of relational named entities, and "mid," which is a unique string of *Freebase* entity attached within a corpus because of Google's *Freebase* as a knowledge base. All five relation types are configured with relations with other entities, including personal names, and sentence data are mixed with sentences with or without representation of relation.

In experiment, 5-fold cross validation was applied to evaluate relation extraction, and the following *Precision* (4), *Recall* (5), and F_1*-Measure* (6) were used as an evaluation method with respect to entity pairs of the knowledge base. Using triples of the experimental data as a knowledge base, experiments were conducted according to each of the evaluation methods.

$$Precision = \frac{Num\ of\ relations\ extacted\ correctly}{Num\ of\ relations\ extacted\ by\ system} \quad (4)$$

$$Recall = \frac{Num\ of\ relations\ extacted\ correctly}{Num\ of\ all\ relations\ corrected} \quad (5)$$

$$F_1 - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

B. Experiment result and analysis

The first experiment compared performances of the existing distant supervision baseline and the proposed system to evaluate the proposed system. The second experiment evaluated clusters that used only semantic similarity calculation without the distant supervision assumption to assess system performance when a knowledge base cannot be used.

Table II shows merged semantic relations according to semantic relation types when WordNet was used with a similarity calculation method. An appropriate threshold value was set via the experiment. The result showed that some semantic relations seemed in error, but expressions with similar meaning can be found such as died, killed, and assassinated. In addition, two sentences "Born in Franksville, Morse attended the schools of Racine County." and "Adam was born in Vitebsk." were extracted into different cluster due to "born" and "Born" but merged into a single cluster later. As such, sentences that have the same verb with different locations can be found here.

TABLE II. Merging Semantic Relations in WordNet Results

Merged Relation	Entity 1	Entity 2	Relations
Place of birth	Person	Location	born, Born, played, brought, making, bringing, grew, left, runs, moving, gave, fly, formed, produced, ...
Place of death	Person	Location	died, killed, assassinated, buried, interred, removed, drowned
Date of birth	Person	Date	born, Born, grew, moved, form, formed, go, went, moving
Institution	Person	Organization	attended, educated, taught, trained, teaching, coached, train, lectured, Coach, lecturing, learned, teaches, inducted
Education degree	Person	Degree	received, obtained

TABLE III. Results of Experiment

Similarity Method	Precision	Recall	F_1-measure
Distant Supervision	0.733	0.163	0.255
Distant Supervision +Wordnet	**0.820**	**0.392**	**0.511**
Distant Supervision +Cosine	0.670	0.266	0.353
Distant Supervision +PMI	0.711	0.240	0.335
Distant Supervision +Collocation	0.797	0.370	0.486

TABLE IV. Results of Experiment Detailed in WordNet

Similarity Method	Relation type	Precision	Recall	F_1-measure
Distant Supervision	Place of birth	0.682	0.192	0.3
	Place of death	0.609	0.107	0.182
	Date of birth	0.733	0.326	0.451
	Institution	0.759	0.116	0.201
	Education Degree	0.884	0.076	0.141
Distant Supervision +Wordnet	Place of birth	0.793	0.334	0.468
	Place of death	0.935	0.224	0.361
	Date of birth	0.786	0.650	0.711
	Institution	0.640	0.282	0.391
	Education Degree	0.946	0.469	0.623

For relation extraction using distant supervision, training data is very important when a classifier is trained. The more the data meets the distant supervision assumption, the better the training result will be. However, the more the data contains error, the worse the performance will be. Here, the baseline method means that training data is tagged in the same way as the distant supervision assumption is applied to each sentence. The experiment results using the distant supervision baseline and similarity calculation methods are shown in Table III.

As shown in Table IV, the experiment results are different depending on the used similarity calculation methods. When similarity was calculated using WordNet, better

performance was revealed than when using only distant supervision but some performance results were comparable or even downgraded. Both Precision and Recall performances were increased in a method using WordNet similarity. The improvement of Precision was due to the effect of exclusion of data that was found to have different semantic relation representations because of low similarity, although the distant supervision assumption was well met. In addition, the reason for the increase in Recall was because the proposed method in this paper used not only sentences that met the distant supervision assumption as training data but also a large number of sentences that belonged to the same group and had a similar format, despite nonconformity to the distant supervision assumption.

Even in the cosine similarity method, performance was degraded because this method compared clusters in a similar way to document comparison, one-by-one but the number of sentences in a cluster was small except for some clusters, and this was why this method had difficulty in similarity calculation. For direct comparison with the cosine similarity method, it would be better to test it over a large amount of data such as big data environment in the future.

Under the PMI method, no improvement on Precision was found. This was because the PMI had difficulty in providing additional semantic information because of a low probability of the same representation in the same sentence due to its characteristic that the same representation was expressed in different sentences in a different manner rather than in the same sentence when the PMI was calculated with respect to represented verbs in a cluster.

The collocation similarity calculation method can circumvent the problem of the PMI method and use peripheral words, which can be an effective measure despite similarity with the PMI. Synonyms, thesaurus, and paraphrase expressions played a similar role in sentences so that co-occurrence words were overlapped, which was extremely helpful in finding paraphrase expressions. This was verified indirectly.

V. Conclusion

The Distant Supervision assumption can be a method that can extract relations as effectively as supervised learning if it is used with a well-refined knowledge base. By using big data, it can show a good effect to save money and maintain performance by removing a cost generated by humans. However, the distant supervision assumption has many potential risk factors as well. Thus, this study proposed a structure that can eliminate error factors of the distant supervision assumption and further use big data.

The experiment results in this study proved that our proposed structure can have positive effects on the distant supervision learning method by semantic elements because the proposed structure can create a large amount of training corpus in a big data environment and distinguish wrong tagging rather than making the distant supervision assumption true through the semantic similarity comparison while using the distant supervision assumption. Furthermore, because the proposed structure can classify using a knowledge base after the training data creation process is complete, it can define classes as many as the number of classes in the knowledge base. When a class in the knowledge base was selected, good performance could be obtained by using a good knowledge base such as *Freebase* when classes were defined using the distant supervision assumption.

References

[1] M. Craven and J. Kumlien, “Constructing Biological Knowledge Bases by Extracting Information from Text Sources,” in Proc. of International Conference on Intelligent System for Molecular Biology, pp. 77-86, 1999. [Online]. Available: http://dl.acm.org/citation.cfm? id=663209

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, “Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge,” in Proc. of SIGMOD, pp. 1247-1250, 2008. doi:10.1145/1376616.1376746

[3] N. Kambhatla, “Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations,” in Proc. of Annual Meeting of the Association for Computational Linguistics, pp. 178-181, 2004. doi:10.3115/1219044.1219066

[4] G. Zhou, J. Su, J. Zhang, M. Zhang, “Exploring various knowledge in relation extraction,” in Proc. of Annual Meeting of the Association for Computational Linguistics, pp. 427-434, 2005. doi:10.3115/1219840.1219893

[5] R. Bunescu, R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction". in Proc. of HLT/EMNLP, pp. 724-731, 2005. doi:10.3115/1220575.1220666

[6] B. Plank, A. Moschitti, "Embedding Semantic Similarity in Tree Kernels for Domain Adaptation of Relation Extraction," in Proc. of Annual Meeting of the Association for Computational Linguistics, pp. 1498-1507, 2013. [Online]. Available: http://www.aclweb.org/anthology/P13-1147

[7] F. Wu, D. Weld, "Open information extraction using Wikipedia,” in Proc. of Annual Meeting of the Association for Computational Linguistics, pp. 118–127, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858694

[8] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, "Open information extraction from the web,” in Proc. of International Joint Conference on Artificial Intelligence, pp. 2670–2676, 2007. doi:10.1145/1409360.1409378

[9] M. Mintz, S. Bills, R. Snow, D. Jurafsky, "Distant supervision for relation extraction without labeled data,” in Proc. of Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003-1011, 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1690287

[10] S. Takamatsu, I. Sato, H. Nakagawa, "Reducing Wrong Labels in Distant Supervision for Relation Extraction,” in Proc. of Annual Meeting of the Association for Computational Linguistics, pp. 721-729, 2012. [Online]. Available: http://dl.acm.org/citation.cfm? id=2390626

[11] I. Augenstein, "Seed Selection for Distantly Supervised Web-Based Relation Extraction,” in Proc. of COLING Workshop on Semantic Web and Information Extraction, pp. 17-24, 2014. [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/I.Augenstein/SWAIE2014-Seed.pdf

[12] X. Zhang, J. Zhang, J. Zeng, J. Yan, Z. Chen, Z. Sui, "Towards Accurate Distant Supervision For Relational Facts Extraction,” in Proc. of Annual Meeting of the Association for Computational Linguistics, pp. 810-815, 2013. [Online]. Available: http://www.aclweb.org/anthology/P13-2141

[13] M. Surdeanu, J. Tibshirani, R. Nallapati, C. Manning, "Multi-instance Multi-label Learning for Relation Extraction,” in Proc. of Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 455-465, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2391003

[14] M. Fan, D. Zhao, Q. Zhou, Z. Liu, T. Zheng, E. Chang, “Distant Supervision for Relation Extraction with Matrix Completion”, in Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 839-849, 2014. [Online]. Available: http://aclweb.org/anthology/P14-1079

[15] T. Nguyen, A. Moschitti, “End-to-end Relation Extraction using Distant Supervision from External Semantic Repositories,” in Proc. of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 277-282, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002794

[16] S. Krause, H. Li, H. Uszkoreit, F, Xu, “Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web,” in Proc. of International Semantic Web Conference, pp. 263-278, 2012.

[Online]. Available: http://www.dfki.de/lt/publication_show.php?id=6420

[17] G. Garrido, A. Penas, B. Cabaleiro, A. Rodrigo, "Temporally Anchored Relation Extraction," in Proc. of Annual Meeting of the Association for Computational Linguistics, pp. 107-116, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390540

[18] M. Surdeanu, D. McClosky, J. Tibshirani, J. Bauer, A. Chang, V. Spitkovsky, C. Manning, "A Simple Distant Supervision Approach for the TAC-KBP Slot Filling Task," in Proc. of Text Analysis Conference, 2010. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.224.7978

[19] Y, Kim. "Automatic Training Corpus Generation Method of Named Entity Recognition using Big Data", Ms. Thesis, Sogang University, 2014. [Online]. Available: http://dcollection.sogang.ac.kr:8089/dcollection/jsp/common/DcLoOrgPer.jsp?sItemId=000000056089

[20] D. Lin. "Extracting Collocations from Text Corpora," Workshop on Computational Terminology, pp. 57-63. 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.7962