JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 28-38, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.3

Minimally Supervised Relation Identification from Wikipedia Articles

Heung-Seon Oh
Korea University of Technology and Education, Cheonan, Korea
E-mail: ohhs@koreatech.ac.kr

Yuchul Jung*
Kumoh National Institute of Technology, Gumi, Korea
E-mail: jyc@kumoh.ac.kr

ABSTRACT
Wikipedia is composed of millions of articles, each of which explains a particular entity with various languages in the real world. Since the articles are contributed and edited by a large population of diverse experts with no specific authority, Wikipedia can be seen as a naturally occurring body of human knowledge. In this paper, we propose a method to automatically identify key entities and relations in Wikipedia articles, which can be used for automatic ontology construction. Compared to previous approaches to entity and relation extraction and/or identification from text, our goal is to capture naturally occurring entities and relations from Wikipedia while minimizing artificiality often introduced at the stages of constructing training and testing data. The titles of the articles and anchored phrases in their text are regarded as entities, and their types are automatically classified with minimal training. We attempt to automatically detect and identify possible relations among the entities based on clustering without training data, as opposed to the relation extraction approach that focuses on improvement of accuracy in selecting one of the several target relations for a given pair of entities. While the relation extraction approach with supervised learning requires a significant amount of annotation efforts for a predefined set of relations, our approach attempts to discover relations as they occur naturally. Unlike other unsupervised relation identification work where evaluation of automatically identified relations is done with the correct relations determined a priori by human judges, we attempted to evaluate appropriateness of the naturally occurring clusters of relations involving person-artifact and person-organization entities and their relation names.

Keywords: relation identification, Wikipedia mining, unsupervised clustering

Open Access

Accepted date: August 08, 2018
Received date: December 08, 2017

***Corresponding Author:** Yuchul Jung
Assistant Professor
Kumoh National Institute of Technology, 61 Daehak-ro, Gumi 39177, Korea
E-mail: jyc@kumoh.ac.kr

1. INTRODUCTION

Wikipedia, the largest online encyclopedia, is composed of millions of articles, each of which explains an entity with various languages in the real world. Since the articles are contributed and edited by a large population of diverse experts with no specific authority, Wikipedia can be seen as a naturally occurring body of human knowledge. This characteristic attracts researchers to focus on mining structured knowledge from Wikipedia.

Relation extraction (RE) often refers to the task of extracting relations between named entities. Most past RE research has focused on development of supervised learning methods for the task of identifying a predefined set of relations from a known corpus, e.g., the ACE corpus. Supervised learning tasks, however, require heavy human annotation efforts to build training data for different domains. To alleviate the problem, semi-supervised methods using a search engine were developed (Etzioni et al., 2005; Pantel & Pennacchiotti, 2006), which start with initial seeds and go through a bootstrapping process using a search engine. Unlike the RE task, recent work on unsupervised relation identification (Hasegawa, Sekine, & Grishman, 2004; Rosenfeld & Feldman, 2006; Rozenfeld & Feldman, 2007; Y. Yan, Okazaki, Matsuo, Yang, & Ishizuka 2009) does not assume a predefined set of target relations, attempting to discover meaningful relations from a given corpus using a clustering algorithm.

As Wikipedia becomes a major knowledge resource, there have been some attempts to extract relations with Wikipedia structural characteristics. A research work (Wu & Weld, 2008) focused on extracting an "infobox" which describes attribute-value pairs of an entity of an article as a way of constructing ontology. A conditional random fields (CRFs) model is automatically trained with sentences related to infobox entries. In Nguyen, Matsuo, and Ishizuka (2007), a system is proposed to extract relations among entity pairs. Rather than using a named entity (NER) tagger to determine the semantic type of an entity, an entity type classifier is trained with features generated from category structures of Wikipedia. Then, relations are extracted with a support vector machines (SVMs) classifier trained by sub-tree features from the dependency structure of entity pairs. Compared to the methods above limited to a set of predefined relations, a method (Y. Yan et al., 2009) was proposed based on unsupervised relation identification framework by incorporating two context types of an entity pair: surface patterns from search results of an entity pair and dependency patterns from parsing the structure of a sentence of an entity pair in Wikipedia. Even though it shows the feasibility of identifying relations in combination with the Web, we thought that considering Wikipedia characteristics to identify relations is much more important.

In this paper, we propose a method to identify meaningful relations from Wikipedia articles with minimal human effort. Our method first detects entity pairs by utilizing the characteristics of Wikipedia articles. Similar to Nguyen et al. (2007), human effort only is required to prepare training data for an entity type classifier. Then, a set of entity pairs not associated in a grammar structure is filtered out. Then, context patterns are generated over sentences with respect to the remaining entity pairs. Based on them, entity pairs are clustered automatically. At last, a cluster label is chosen by selecting a representative word for each cluster. Experimental results show that our method produces many relation clusters with high precision. In previous work (Nguyen et al., 2007; Y. Yan et al., 2009), analysis of utilizing the characteristics of Wikipedia was not reported in detail even though the importance of the characteristics is not addressed. This paper reports our deep investigation.

The rest of this paper is organized as follows. Section 2 briefly introduces relevant research. The details of our method are described in Section 3. Section 4 delivers experimental results. Finally, we conclude in Section 5 with a suggestion for future work.

2. RELATED WORK

Wikipedia has been utilized for other purposes. Semantic relatedness (Gabrilovich & Markovitch, 2007; Strube & Ponzetto, 2006) is measuring the relatedness of two words or phrases utilizing characteristics such as the unique names of the articles and category hierarchy. Text classification (Gabrilovich & Markovitch, 2006) also utilizes the unique names of Wikipedia articles. Rather than using a bag of words approach, it utilizes the names of Wikipedia articles as semantic concepts for input text. When two input texts are entered, they are mapped to articles including each text and get the names of the articles as semantic concepts. The concepts are used as features for text categorization. Wikipedia also was used in taxonomy or ontology generation (Strube & Ponzetto, 2006; Wu & Weld, 2008). Due to the various usages of Wikipedia, the tasks of extracting entities and relations from Wikipedia are quite meaningful.

There have been some attempts to extract entities and relations from Wikipedia. One research work (Culotta,

McCallum, & Betz, 2006) regards RE as a sequential labeling task like NER and applies a CRFs model with conventional words and patterns as features for learning a classifier. In Nguyen et al. (2007) an entity detector and SVMs classifier were built using the characteristics of Wikipedia articles. Then, relations among the detected entities were determined by using another SVMs classifier trained with sub-trees mined from the syntactic structure of text. Unlike our approach, these approaches restrict target relations and require a significant amount of human labor for building the training data. KYLIN (Wu & Weld, 2007) automatically generates training data using infoboxes of Wikipedia articles to learn a CRFs model and extracts attribute-value pairs from the articles that have incomplete or no infoboxes.

Open information extraction (OpenIE) is a research area aiming to extract a large set of verb-based triples (or propositions) from text without restrictions of target entities and relations. Reverb (Fader, Soderland, & Etzioni, 2011) and ClauseIE (Corro & Gemulla, 2013) are representative projects to pursue OpenIE. Due to the no restrictions, OpenIE systems tried to consider all possible entities and relations in text of interest and thus produces many meaningless extractions. Unlike OpenIE, we are interested in somewhat normalized entities and relations existing in Wikipedia.

For the task of unsupervised relation identification, a research work (Hasegawa et al., 2004) shows a successful result of applying clustering to relation discovery from large corpora. It detects named entities using a NER tagger and considers entity pairs that often co-occur in a corpus for relation discovery. Entity pairs with intervening words between them are clustered using a hierarchical clustering technique. For each cluster, a representative word is chosen as the relation name based on word frequency. Instead of using intervening words, other systems (Rosenfeld & Feldman, 2006; Rozenfeld & Feldman, 2007) adopted a context pattern extraction and selection methods that uses dynamic programming and an entropy-based measure among the extracted patterns, respectively. The relation identification method in our system resembles the aforementioned method but with some unique technical details for a different resource, namely, Wikipedia.

In recent work (D. Zeng, Liu, Lai, Zhou, & Zhao, 2014), neural networks are employed to train an extraction model. D. Zeng et al. (2014) utilized the convolutional neural network to automatically extract features that are not dependant on traditional natural language processing tools and evade the error propagation problem. Although other approaches based on deep learning adopted long short-term memory networks along the shortest dependency path (X. Yan et al., 2015) and proposed an attention mechanism with bidirectional long short-term memory networks (Zhou et al., 2016), all of these models require sufficient training data and time to generate a high-performing model.

To alleviate the difficulties of producing training examples for RE, distant supervision has been used (Craven & Kumlien, 1999; Mintz, Bills, Snow, & Jurafsky, 2009). There exist two major research directions for the distant supervision. One direction is to use it for directly enriching knowledge bases from unstructured text, as well as leveraging the knowledge bases to generate the distant supervision labels (Poon, Toutanova, & Quirk, 2015; Parikh, Poon, & Toutanova, 2015). The other direction, so called Socratic learning (Varma et al., 2016), uses the differences in the predictions of the generative model to reduce the noise in distant supervision labels. Meanwhile, those approaches require multiple sources of weak supervision. More recently, a reinforcement learning approach was proposed to conduct large scale RE by learning a sentence relation extractor with distant supervised datasets (X. Zeng, He, Liu, & Zhao, 2018).

3. PROPOSED METHOD

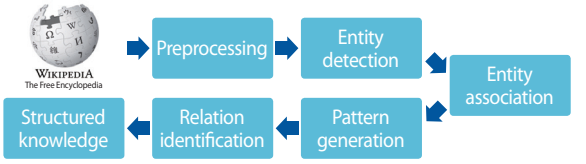

Fig. 1. Overview of proposed method.

Fig. 1 shows the overview of our method. From Wikipedia articles as input, structured knowledge is identified with minimal human effort. At first, preprocessing is performed, such as tokenization, Part-of Speech tagging, and chunking to Wikipedia articles. For each sentence, entities are detected and associated to make entity pairs. Discriminative patterns for entity pairs are retained. Entity pairs are clustered based on the patterns with hierarchical clustering method. Then, for each cluster, a representative word is selected as a name of the cluster.

3.1. Preprocessing

Several preprocessing stages are performed on Wikipedia articles. We first retain the raw text of an article by filtering out markup tags. Then, several miscellaneous parts not related to the main text such as *See Also* and *References* are

discarded. The remaining text parts of the articles undergo tokenization, sentence splitting, Part-of Speech tagging, and chunking steps in turn via OpenNLP tools.[1]

To ensure that a sufficient amount of contextual information exists surrounding entities, we discarded sentences having less than five words, and articles consisting of less than 25 sentences. Sentences with more than 30 words were also discarded to avoid potential errors due to the complexity involved in sentence processing.

3.2. Entity Detection

In Culotta et al. (2006), two types of entities are defined in Wikipedia articles: a principal entity and secondary entity. A principal entity refers to an instance of the name (title) of the article which is being described. A secondary entity refers to mentioned entities anchored in the same article which is linked to another Wikipedia article. A principal entity is often expressed in a different way with an anaphor. This is a natural phenomenon of English. For example, "Bruce Willis," a famous movie star, can be mentioned with "Willis," "he," or "an American actor" in the corresponding article. Definitely, we may miss many mentions of a principal entity without considering anaphors. There are various methods to resolve anaphora and co-references (Sukthanker, Poria, Cambria, & Thirunavukarasu, 2018). We adopted the heuristic method in Nguyen et al. (2007) for resolving anaphors referring to principal entities. Secondary entities linked to other Wikipedia articles are identified in a straightforward manner as they are tagged as such. Entities ending with a proper noun are only considered since our current focus is on named entities. The above step results in sentences with a principal and secondary entity pair.

To retain meaningful relations, the semantic classes of entities should be considered. For example, a chairman relation only occurs between person and organization. In our work, four semantic classes of entities are considered: person, organization, location, and artifact. An article does not belong to any of four semantic classes because they do not cover all Wikipedia articles. For that reason, we add other types for undefined classes.

As each entity corresponds to a Wikipedia article, entity classification can be regarded as text classification aiming at classifying an article to one of five classes. Unlike a common text classification, we assumed that all parts of an article are not effective to classify among five semantic classes. Similar to Nguyen et al. (2007), the SVMs classifier is trained with five features incorporating Wikipedia's structural characteristics: 1) category feature (categories collected by tracking back from the article up to k parent levels of the Wikipedia category hierarchy), 2) category term feature (the terms in the category feature), 3) category headword feature (the headwords of categories in the category feature), 4) first sentence term feature (terms in the first sentence in the article), and 5) title term feature (terms consisting of the article title). In this step, human effort is required to prepare an annotated dataset. Fortunately, this is cheap and easy because our task is just to assign a semantic class to an article, not a label sequence of a word sequence, for common NER tasks.

3.3. Entity Association

A major goal of our research is to identify relations between principal and secondary entities in a Wikipedia article. To satisfy the goal, we should find potentially useful entity pairs that can have a certain relation. Two approaches are possible: based on co-occurrence or a grammatical relation between two entities. The first approach as used in Hasegawa et al. (2004) selects entity pairs that occur more frequently than a threshold. A pair of entities that occur together very rarely would not possess a relation of sufficient interest. The second approach selects entity pairs involved in a grammatical relation, like a subject-object or object-subject relation, as in Shinyama and Sekine (2006).

Unlike more frequently used data for relation extraction, such as news data, however, there are few co-occurring entity pairs in Wikipedia because of the nature of

(a) Entity pair information with corresponding sentences

ID	First Entity	Second Entity
1	(0, 0, P, PER, He/PRP)	(2, 3, S, PER, Dan/NNP Rather.NNP)
2	(0, 0, P, PER, He/PRP)	(6, 7, S, ORG, Planet/NNP Hollywood/NNP)

ID	Sentences
1	He/PRP interviewed/VBD Dan/NNP Rather/NNP in/IN what/WP he/PRP would/MD later/PB call/VB the/DT most/RBS serious/JJ conversation/NN of/IN my/PRP$ entire/NN life/NN *J*.
2	He/PRP is/VBZ also/RB a/DT co-founder/NN of/IN Planet/NNP Hollywood/NNP *J*.

(b) Slot-marked sentences

ID	Sentences
1	<PER>/ENT interviwed/VBD <PER>/ENT in/IN what/WP he/PRP would/MD later/RB call/VB the/DT most/RBS serious/JJ conversation/NN of/IN my/PRP$ entire/NN life/NN *J*.
2	<PER>/ENT is/VBZ also/RB a/DT co-founder/NN of/IN <ORG>/ENT *J*.

Fig. 2. Entity pair information for the article on "Bruce Willis" (a) and sentences after slot-marking (b).

[1] https://opennlp.apache.org/

encyclopedia articles. For that reason, we parsed sentences and retained predicate-argument structures. Based on the structure, sentences with entity pair matched to subject-object pair are assumed to have a relation. Fig. 2 shows an example "Bruce Willis" article. The entry (0, 0, P, PER, He/PRP) indicates the start token, end token, principal entity, person type, and the entity text, respectively.

For further processing, entities in sentences are generalized by being slot-marked with a corresponding entity type and ENT indicating an entity tag. In addition, numbers are normalized to "#NUM#". This generalization process makes it easier to find common patterns for clustering. An example for a slot-marked sentence is shown in Fig. 2(b).

3.4. Pattern Generation

To identify relations, each entity pair is encoded as a feature vector representation. A feature vector should consist of discriminative features and values. To satisfy two conditions, feature vectors are constructed through a pattern extraction and selection (Fradkin & Mörchen, 2015).

The aim of pattern extraction is to provide necessary data for clustering entity pairs. In order to provide sufficient context information of entities, we applied Smith-Waterman (SW) algorithm (Smith & Waterman, 1981), which is one of the dynamic programming methods for a local alignment of molecular subsequences, for context pattern extraction.

The SW algorithm starts with constructing a score matrix D for two different input sentences using the scoring scheme shown below. The two input sentences are represented as $s = s_0 s_1 \cdots s_i$ and $t = t_0 t_1 \cdots t_j$ where s_i and t_j indicates i-th and j-the words in the two input sentences, respectively.

$$D(i,j) = \max \begin{pmatrix} 0 \\ D(i-1,j-1) + D(s_i, t_j) \\ D(i-1,j) - gap \\ D(i,j-1) - gap \end{pmatrix} \quad (1)$$

Here D(i, j) is a cost function for i-th and j-th words and gap is a penalty cost for a gap. We set gap to 1 and defined the cost function below.

$$D(i,j) - \begin{cases} 2 & \text{if } s_i = t_j \\ -1 & \text{if } s_i \neq t_j \end{cases} \quad (2)$$

Initially, all positions of the score matrix are initialized with 0. By comparing s_i and t_j, the score matrix is filled with D(i, j). After constructing the score matrix, backtracking is carried out for finding the best local alignment starting from the position assigned a maximum score on the matrix following the policies in turn.

$$D(i,j) - \begin{cases} D(i-1,j-1) & \text{if } D(i,j) = D(i-1,j-1) + D(s_i, t_j) \\ D(i-1,j) & \text{if } D(i,j) = D(i-1,j) - gap \\ D(i,j-1) & \text{if } D(i,j) = D(i,j-1) - gap \end{cases} \quad (3)$$

Fig. 3 shows an example for computing the alignment matrix and the resulting alignment between two input sentences.

An alignment is converted to a pattern after replacing mismatching and similar words with a wild card character that allows for any word sequence. In the example, the alignment is converted to a pattern "<ORG> * located in * <LOC>".

(a) Computing an alignment matrix

	<ORG>	wes	located	in	near	<LOC>	.
<ORG>	2	1	0	0	0	0	0
is	1	1	0	0	0	0	0
located	0	0	3	2	1	0	0
in	0	0	2	5	4	3	2
<LOC>	0	0	1	4	4	6	5
.	0	0	0	3	3	5	8

(b) An alignment between two sentences

<ORG>	is	located	in	GAP	<LOC>	.
|	:	|	|	.	|	|
<ORG>	wes	0	3	3	5	5

Symbol	Meaning
|	Match
.	Mismatch
:	similar

Fig. 3. Example of computing an alignment matrix (a) with the resulting alignment (b).

Even though pattern extraction aims at reflecting common contextual information of entity pairs, not all of the patterns are helpful for identifying relations. Many patterns are not discriminative because they are too specific or too general to certain contexts. In the clustering phase, such patterns may introduce noise and result in unexpected entity pair clusters. As such, selecting patterns with sufficient entity revealing contextual information is critical.

There are several feature selection methods such as information gain and x^2 that work with labeled data (Forman, 2003). However, they are not applicable because we do not have labeled data for relations. For that reason, an unsupervised feature selection method is adopted for selecting useful patterns (Jinxiu, Donghong, Lim, & Zhengyu, 2005; Rosenfeld & Feldman, 2007). The intuition behind the method is that good clustering features should

improve the separability of the dataset, making points that are close together still closer, and points that are far from each other still farther apart.

Let $C = \{c_0, c_1, \ldots, c_n\}$ be a set of examples where an example consists of patterns as features. Then, cosine similarity between two examples is defined:

$$S(c_i, c_j) = S_{ij} = \frac{c_i \cdot c_j}{|c_i||c_j|} \quad (4)$$

Using the similarity, scoring function for a feature f is defined:

$$Score(f) = E - E_{-f} \quad (5)$$

Where

$$E = -\sum_{c_i, c_j \in C} S_{ij} \log S_{ij} + (1 - S_{ij}) \log(1 - S_{ij}) \quad (6)$$

$$E_{-f} = -\sum_{c_i, c_j \in C} S_{ij}^{f} \log S_{ij}^{f} + (1 - S_{ij}^{f}) \log(1 - S_{ij}^{f}) \quad (7)$$

and S_{ij}^{f} is the similarity between c_i and c_j after removing the feature f.

Performing the feature selection for full feature space over all examples is very time-consuming. To reduce the feature space with retaining patterns directly related to entities, we discard patterns which do not have entity slots and content words such as noun, verb, and adjective before feature selection. For example, "* located in *" is discarded because no entity slot occurs.

3.5. Relation Identification

Our goal is to discover relations from all entity pairs represented as a set of discriminative patterns. For that reason, a hierarchical agglomerative clustering (HAC) algorithm which is not concerned with the number of clusters in advance is a natural choice. As reported in Rosenfeld and Feldman (2007), we opted for single link HAC because it outperforms average and complete link HACs for relation identification tasks.

In single link HAC, initially, each of the data points is regarded as a single cluster. When the similarity distance of two clusters is within a threshold, two clusters merge. As a result, determining the threshold affects the clustering results. In our case, we utilized cosine similarity and set the threshold to 0.3. Since clusters without a sufficient number of instances cannot have a representative for the identified relation, those with less than five instances were not considered for further processing.

Since entity pairs are clustered based on the similarities of context patterns, we can assume that instances in each cluster have a common meaning for the context patterns, i.e., a relation between entities in our case. Instead of classifying the meaning to one of the existing relation names as in RE tasks, we opted for naming it with a representative word found in the cluster. The terms between the two entities in a cluster are candidates and evaluated with the TF*IDF scheme where TF is the term frequency in the cluster and IDF is the inverse document frequency of the term over entity paired sentences. The identified relations are shown in the last of this paper.

4. EXPERIMENTS

For experiments, we downloaded English Wikipedia articles and randomly selected a total of 32,355 articles after filtering, where an article was filtered if it did not represent a real-world entity. For example, entity Forrest Gump was discarded because he is not an actual person but the main character of a movie, while Tom Hanks, an actor who played the character, was kept because he is a real world entity. After going through entity detection and association explained in subsections 3.2 and 3.3, 103,526 sentences with principal and secondary entity pairs were retained. For example, let us see the sentence "Hanks has collaborated with film director Steven Spielberg on five films to date." Hanks is a principal entity while Steven Spielberg, a famous movie director, is a secondary entity in the article "Tom Hanks."

To assign the semantic classes of each entity, we built an entity classifier with LIBSVM (Chang & Lin, 2011). 4,123 and 415 articles were manually annotated and tested. Table 1 shows the results of the entity classifier. We obtained the best result performance when all features such as 1) category feature, 2) category term feature, 3) category headword feature, 4) first sentence term feature, and 5) title term feature were used with up to four parents in category structure.

Table 1. Performance of entity classifier

Features	Parent levels	Accuracy
All features	3	0.8364
	4	0.8571
	5	0.8475
	6	0.8356

Table 2. Results of pattern extraction and selection (# of instances)

Domain	Sentence	Extracted pattern	Selected pattern
PER-ART	4,567	14,856	3,782
PER-ORG	6,703	29,800	9,603

Table 3. Results of clustering with entity pairs

Domain	Relation cluster (Entity Pair)	Garbage cluster (Entity Pair)
PER-ART	115 (1,549)	1,548 (2,229)
PER-ORG	160 (2,180)	3,304 (4,015)

Table 4. Performances of anaphor identification and entity classification

Criterion	Total	Correct	Precision
Anaphor (PER-ART)	1,549	1,467	0.947
Anaphor (PER-ORG)	2,180	2,125	0.975
Entity (PER-ART)	3,098	2,382	0.769
Entity (PER-ORG)	4360	4280	0.982

To analyze the results in detail, we focused on two domains, person-organization (PER-ORG) and person-artifact (PER-ART). Table 2 shows simple statistics resulting from pattern extraction and selection for the two different cases. It can be seen that the number of surviving patterns after the selection process is only one third of the extracted patterns.

For clustering of entity pairs, we utilized LingPipe,[2] freely usable natural language tools, for single link HAC. Clusters that contain less than five entity pairs are considered a garbage cluster. Table 3 shows the results of clustering. In the case of PER-ART, for example, a total of 1,549 entity pairs form 115 relation clusters, indicating that 1,549 entity-relation-entity triples with 115 relations can be generated.

In entity detection, a heuristic method is adapted for identifying anaphors of principal entities. The effects of anaphor identification should be investigated because many entity pairs include anaphors and are processed further.

Table 4 shows the performances of anaphor identification and entity classification. It shows promising results in both domains. However, the precision of entity classification in the PER-ART domain is surprisingly lower than that of the PER-ORG domain, indicating that entity classification for

[2] http://alias-i.com/lingpipe/

Table 5. Precisions on PER-ART domain

Case	Total	Correct	Precision
1	1,549	1,093	0.706
2	1,467	1,059	0.722
3	836	626	0.749
4	798	604	0.757

Table 6. Precisions on PER-ORG domain

Case	Total	Correct	Precision
1	2,180	1,841	0.844
2	2,125	1,796	0.845
3	2,099	1,782	0.849
4	2,056	1,745	0.849

ART is more difficult than that of ORG. We have found two reasons resulting in the performance drop. The first is that entity classification is conducted for each article, not for each sentence. As a result, every entity receives the same entity type regardless of the context of an entity pair in a sentence. For example, in the following two sentences, Singapore General Hospital is supposed to have two different entity types: organization for the first and artifact for the second.

1. Ratnam began his career as a houseman at the Singapore General Hospital in 1959.
2. Singapore General Hospital was built in 1920.

The second reason is the insufficient coverage of training data. For example, many historical war names such as American Civil War are classified as artifacts. However, they should be classified as other categories and filtered out for further processing. It turns out that those incorrectly classified entities share the same category hierarchy information from Wikipedia, which is a key feature for our classifier, with those correctly classified. We evaluated the appropriateness between an entity pair and a relation by determining whether or not a relation is a representative word for an entity pair. For that, an entity pair and a relation are represented as a relation triple like entity-relation-entity. As a result, a precision indicates the overall appropriateness with respect to all of the relation triples. In order to avoid biased subjectivity, we counted a relation triple for precision when two evaluators (i.e., two authors of this paper) both agree with a relation triple as being appropriate.

Tables 5 and 6 show the results for each domain. To

Table 7. Example of cluster name errors

Relation	Example
German	Hattie McDaniel was the first performer of African descent to win an Academy Award.
enrolled	Bruce R. McConkie enrolled in Army ROTC while at the University of Utah.

analyze the effect of erroneous results of entity detection, we conducted four different evaluations for each domain. Case 1 includes all of the incorrect results from anaphor identification and entity classification. Case 2 excludes the incorrect results from anaphor identification. Case 3 excludes the incorrect results from entity classification. Case 4 excludes all of the incorrect results from anaphor identification and entity classification. The results show that excluding the incorrect results in the earlier phases improves the precision .051 and 0.005 in PER-ART and PER-ORG domains, respectively.

Considering only case 4 of both domains, we found two error types. The first error type shown in Table 7 is that the identified relation is not appropriate to represent the relation among entities. The second error type is caused by incorrect subject-object entity pairing. In the second example of Table 7, University of Utah is not an object of Bruce R. McConkie. This error type entirely depends on the results of parsing predicate-argument structure.

5. CONCLUSION

In this paper, we presented a method that identifies naturally occurring relations between entities in Wikipedia articles with an aim to minimize human annotation efforts. The manual annotations are required to construct training data for an entity classifier in general. However, the efforts should be minimized because it is a simple task of assigning a class to a Wikipedia article. Using the entity classifier, entity pairs which may have a meaningful relation are kept for relation identification. Relations are identified in an unsupervised way based on hierarchical clustering and pattern generation and selection. Our experimental results showed promising results for both entity classification and relation identification. From the analysis of experiments, we found that error propagation from entity classifier and heuristic anaphora detection is a critical issue for improving performance, but hard to avoid since our method heavily relies on unsupervised learning.

As Wikipedia grows and evolves via the contributions of the general public, this kind of automatic identification of relations among key entities that reflect the real world would be very useful for a variety of applications such as ontology and knowledge base construction, guided searching and browsing, and question answering. More specifically, in aspects of ontology construction, our proposed methods can be effectively used for building core (basic) ontology of specific domains. After that, the core ontology can be populated further by combining with domain-specific patterns, knowledge-based approaches, and other state-of-the-art supervised/unsupervised approaches.

Our future work includes the following extensions: expansion of the entity type pairs, more thorough and larger scale evaluation of the relation identification task, and more direct evaluation of the value of the entity and relation identification for ontology construction.

ACKNOWLEDGMENTS

This work was supported by the research fund for a newly appointed professor of Korea University of Technology & Education in 2018. This work was also supported b y the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF -2018R1C1B5031408).

REFERENCES

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Corro, L. D., & Gemulla, R. (2013). ClausIE: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 355-365). New York: ACM.

Craven, M., & Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* (pp. 77-86). Menlo Park: AAAI Press.

Culotta, A., McCallum, A., & Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (pp. 296-303). Stroudsburg: Association for Computational Linguistics.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S.,…Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1), 91-134.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1535-1545). Stroudsburg: Association for Computational Linguistics.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 1289-1305.

Fradkin, D., & Mörchen, F. (2015). Mining sequential patterns for classification. *Knowledge and Information Systems*, 45(3), 731-749.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1301-1306). Menlo Park: AAAI Press.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence* (pp. 1606-1611). San Francisco: Morgan Kaufmann Publishers.

Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (pp. 415-422). Stroudsburg: Association for Computational Linguistics.

Jinxiu, C., Donghong, J., Lim, T. C., & Zhengyu, N. (2005). Unsupervised feature selection for relation extraction. In R. Dale, K. F. Wong, J. Su, & O.Y. Kwong (Eds.), *Natural Language Processing: IJCNLP 2005* (pp. 390-401). Berlin: Springer.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 1003-1011). Stroudsburg: Association for Computational Linguistics.

Nguyen, D. P. T., Matsuo, Y., & Ishizuka, M. (2007). Relation extraction from Wikipedia using subtree mining. In *Proceedings of the 22nd National Conference on Artificial Intelligence* (pp. 1414-1420). Menlo Park: AAAI Press.

Pantel, P., & Pennacchiotti, M. (2006). Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 113-120). Stroudsburg: Association for Computational Linguistics.

Parikh, A. P., Poon, H., & Toutanova, K. (2015). Grounded semantic parsing for complex knowledge extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 756-766). Stroudsburg: Association for Computational Linguistics.

Poon, H., Toutanova, K., & Quirk, C. (2015). Distant supervision for cancer pathway extraction from text. In *Pacific Symposium on Biocomputing Co-Chairs* (pp. 120-131). Singapore: World Scientific.

Rozenfeld, B., & Feldman, R. (2006). High-performance unsupervised relation extraction from large corpora. In *Proceedings of Sixth International Conference on Data Mining* (ICDM'06) (pp. 1032-1037). Piscataway: IEEE.

Rosenfeld, B., & Feldman, R. (2007). Clustering for unsupervised relation identification. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (pp. 411-418). New York: Association for Computing Machinery.

Shinyama, Y., & Sekine, S. (2006). Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (pp. 304-311). Stroudsburg: Association for Computational Linguistics.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195-197.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1419-1424). Menlo Park: AAAI Press.

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2018). *Anaphora and coreference resolution: A review*. Retrieved September 2, 2018 from https://arxiv.org/pdf/1805.11824.pdf.

Varma, P., He, B., Iter, D., Xu, P., Yu, R., De Sa, C., & Ré,

C. (2016). *Socratic learning: Augmenting generative models to incorporate latent subsets in training data*. Retrieved September 2, 2018 from https://arxiv.org/abs/1610.08123.

Wu, F., & Weld, D. S. (2007). Autonomously semantifying Wikipedia. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (pp. 41-50). New York: Association for Computing Machinery.

Wu, F., & Weld, D. S. (2008). Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 634-644). New York: Association for Computing Machinery.

Yan, X., Mou, L., Li, G., Chen, Y., Peng, H., & Jin, Z. (2015). Classifying relations via long short term memory networks along shortest dependency path. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1785-1794). Stroudsburg: Association for Computational Linguistics.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., & Ishizuka, M. (2009). Unsupervised relation extraction by mining Wikipedia texts using information from the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 1021-1029). Stroudsburg: Association for Computational Linguistics.

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics* (pp. 2335-2344). Sheffield: International Committee on Computational Linguistics.

Zeng, X., He, S., Liu, K., & Zhao, J. (2018). Large scaled relation extraction with reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 5658-5665). Palo Alto: Association for the Advancement of Artificial Intelligence.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 207-212). Stroudsburg: Association for Computational Linguistics.

APPENDIX. SAMPLE CLUSTER DETAILS

Table A1. Sample cluster details in PER-ART domain

Relation	Entity pair	Context pattern	Entity paired sentence	True	False
appear	Tyra Banks- Felicity Tamera Mowry- Smart Guy	<PER> appeared on <ART> <PER> appeared in	<PER> also appeared on <ART>. <PER> appeared in <ART >.	35	2
role	Chris Elliott - Cabin Boy Sylvester McCoy- The Cabaret of Dr Caligari	<PER> had * role in * <ART> <PER> played * role of *	<PER> had title role in <ART>. <PER> played the role of Snuff in the macabre BBC Radio 4 comedy series <ART>.	24	4
performed	Kellie Pickler- Red High Heels Alan Autry:Autry- Rudolph the Red Nosed Reindeer	<PER> performed * <ART> <PER> performed * of <ART>	<PER> performed live <ART> <PER> performed his rendition of <ART>	13	1
won	Philip K. Dick- The Man in the High Castle Robert Fuller- Golden Boot Award	<PER> won * <ART> for * in In #NUM# * <PER> won * <ART>	In 1963, <PER> won the Hugo Award for <ART >. In 1989, <PER> won the <ART>.	32	5

Table A2. Sample cluster details in PER-ORG domain

Relation	Entity pair	Context pattern	Entity paired sentence	True	False
educated	Haldane - Dragon School Dick McCreery- Eton College	<PER> was educated at <ORG> * College <PER> was educated at <ORG> * , * and	<PER> was educated at <ORG>, Eton College and at New College, Oxford. <PER> was educated at <ORG > .	48	0
professor	Haushofer - University of Munich Von Laue - University of Zurich	<PER> * professor * at <PER> * professor of * at * <ORG>	In 1919, <PER> would become professor of geography at the <ORG>. <PER> became professor of physics at the <ORG> in 1912.	42	0
attended	Brookings - Bowdoin College Hicks - University of Houston	<PER> attended <ORG> * , <PER> * attended * <ORG>	<PER> attended <ORG> in Brunswick. <PER> also attended the <ORG> for a short time.	140	3
member	Vance Plauche - American Legion Merlin Olsen - Phi Beta Kappa	<PER> was * member of <ORG> <PER> * member of * <ORG> and * in	<PER> was also a member of the <ORG>. <PER> is a member of Sigma Chi fraternity and <ORG > and was a letterman in football as a defensive tackle.	140	0