Journal of Theoretical and Applied Information Technology
20th April 2015. Vol.74 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NOVEL RULE BASED APPROACH FOR ENTITY RELATIONS EXTRACTION

[1]MUJIONO SADIKIN, [2]ITO WASITO
[1]Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia
[2]Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
E-mail: [1]mujiono.sadikin@mercubuana.ac.id , [2]ito.wasito@cs.ui.ac.id

ABSTRACT

There are always new challenges in the extraction of object (entity) relations contained in unstructured or semi structured text documents found in the Internet, due to the volume of the documents, the evolution of the text language, and the fast Internet growth. In this paper, the authors present the description and the experimental results of a novel role based approach in mining the entities and its relations. The proposed method defines a new concept of entity relationship which treat entities relation as the relation of the main object and its supporting object. The relation between these objects are extracted through pattern learning process that utilize the Indonesia WordNet as an external knowledge. Based on the performance evaluation of the proposed method, it can be confirmed that it is feasible to apply the method in the area. The feasibility of the method is measured by the accuracy of the extraction process in 10 experiments. The average F-score values for the experiments are 0.895 and 0.795 in main object extraction and supporting object extraction respectively.

Keywords: *object extraction relation, object, object interaction, pattern learning, tuple scoring*

1. INTRODUCTION

Text narration in a document is actually describing the entities and their semantic relation. There is also a possibility that the entities in a certain document semantically interact with other entities in other document and perform an entities interaction. This paper, as an example, contains some entities such as authors, topic, method, data set, or references. In this case, several relations can be built between authors and topic, topics and methods, or method and the references. The relation between author and topics can be "to discuss" or "to study". Another example is text narration contained in a drug label document. In a drug label document, several entities are contained in the text narration. They are the drug name, the drug component, the indications, and the contraindication or the side effects. The semantic relation between drug name and drug component is "this drug name contains the chemical component". The drug name in a certain drug label document can also interact with other drug names on other drug label documents, as they have common or similar component. Similarly, authors of a certain paper semantically related (or interact) with other authors of another paper because of, for example, the similarity of the topics discussed.

1.1. Related Study

Even though numerous experts have done works concerning and various approaches have been proposed in the study of entity relation extraction and or interaction, there are still open wide range of challenges can be explored in this research area [1]. Some of the reasons why the relation extraction task is very challenging are the great diversity of various domain corpora and the limited number of labeled training set [2]. The challenge of the entities and entity relation extraction is even greater when the document in question comes from the World Wide Web, since there is no limit to the growth of a web document and also the unique characteristic of such document [3]. In essence, new techniques and approaches are always required to extract the entities and their relationships or their interaction because of several reasons such as the day-by-day growth of the corpus volume, the differences of the languages used in the document, or the evolution of each of the language used. As presented in ACE workshop [4], there are several relation extraction approaches that have been proposed within the framework of the various technique. Most relation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extraction frameworks which are developed so far can be classified into two methods, supervised and semi supervised [5]. Supervised method works are based on kernel, feature, and concern; while semi supervised methods include rule based and pattern based method.

Several supervised methods which are based on the feature are presented in [6], [2], [7], [8], [9], [10], while ones which are based on kernel are proposed in [11],[12], [13], [14], [15], [16], and [17]. Some of these methods are described briefly in this section. The supervised relation extraction method that is proposed in [7] is used to extract the Chinese word relation. The author constructs five main features to perform the method described in the framework. Those features are Entity Type and Subtype, Head Noun, Position Feature, POS Tag, and Omni Word. The more complex relation extraction is explored with the utilization of the hyperlink of the Wikipedia web page to extract the hyponym relation between entities, as is proposed in [2]. In the study, the author built 13 domains feature vector of three node types that are stated as a motif of a Wikipedia Article Graph (WAGs). The method uses heuristic approach in labeling the training set of the domain that is based on the extraction of the structural information, and base on this training set it performed the classification model to discover hyponym relations. Z. Xu et al [6], proposed a Temporal Semantic Relation (TSR) extraction methods which is based on the integration of features collected from the dictionary, IEEE, and Renlifang. The main step of this feature based method are: *look up the Google TSR repository, generate Lexical Semantic Pattern (LSP), generate the entities connection of the pair of entities, generate context sentences, generate context graphs, generate context community, and optimize the LSP*.

A good explanation about the using of kernel to extract entity relation can be referred in [12]. This paper presents the modification of kernel function that applied to syntactic sources. This paper presents the modification of kernel function that is applied to syntactic sources. The kernel modification incorporates information tokenization, parsing and dependency analysis. For this purpose, four kernels are performed, they are the argument kernel, then bigram kernel, the link sequence kernel and the dependency path kernel. The utilization of kernel which is applied to the syntactic parse tree is presented in [13] and [14]. Paper [17] described the application of Convolution Tree Kernel to the tree to compute the similarities between relations, while in the last paper [14], the simple Kernel is used to represents the dependency tree. In this paper, the author utilizes the efficiency of Kernel function computation to calculate the common features of two or more relations.

Another approach for relation extraction purposes is rule based learning [18]. Some of the approaches are proposed in [19], [20], [21], [22], and [23]. One of the rules based learning methods is a bootstrapping approach. The first generation of bootstrapping approach is Snowball [20]. In this study, the author applied bootstrapping technique for extraction of binary relations, such as Organization-Location, e.g., between Microsoft and Redmond, WA. Thellen M and Riloff E [22] proposed bootstrapping method to infer semantic lexicon of new words. The bootstrapping method was utilized to perform new patterns to identify the new word category. Pattern learning with bootstrapping approach is also studied by W. Lin et al [22] to extract the names entities of a certain domain. In this case, the author applied the pattern learning algorithm to disease and location category. Another pattern learning approach to extract information of specific domain was proposed in [23]. The other most recent study related to bootstrapping is proposed by Liu, Ting and Tomek S. [24]. In their study, the bootstrapping method extracts events and its relation from text, based on the resulted pattern from the learning process. These learning processes include two mechanisms: learning through pattern mutation and learning by exploiting structural duality. Event information viewed as multiple faces, which is extracted from news, is also studied with bootstrapping approach. This kind of learning pattern – bootstrapping application was published in [25].

1.2. Overview of Our Approach

In this study, we propose a new concept of entity relationship contained in a document. We introduce the concept of entities as main object (MO) and entities as supporting object (SO). The existence of supporting object is dependent on the main object, but not in the other way around. Therefore, there cannot be a supporting object if there is no main object, but there can be main objects, even though they do not have any supporting object. In a drug label document, for example, entities acting as the main object are drug names and the supporting objects can be its components, its indication, or its contraindication. If the document is the product brochures, the main object is the product name while its supporting object can be its dimension, its specification, its product components, or its price.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This paper is a part of a series of researches and its main objective is to predict object interaction in the corpora discussed in the previous study [26]. As a single phase of the series of researches, the goal of the study in this paper is to build a model that can be applied to the main object (MO) and supporting object (SO) relation extraction. In the study, we do the experiments to extract drug-name – drug-component from various drug label documents which are grabbed from several web sites. For relation extraction purposes, the authors use pattern learning that is based on regular text expression that surrounds the targeted entities. To generate the pattern from the training data that will be applied to the testing data, it uses the Bahasa Indonesia (Indonesian Language) WordNet as an external knowledge. More detail regarding the proposed method is explained in section 2.

1.3. Contribution and Organization

The main contributions of this study are:

1) A new concept of relations between entity act as main object and other entities act as supporting objects.
2) A relation extraction method that is independent to the structure of human Natural Language.
3) In framework proposed by the authors, there is no need to do the preprocessing that is commonly applied to NLP text, such as sentence parsing, stop word removal, or POS Tagging.
4) There is no need to provide prior knowledge in the form of ontology or hierarchical knowledge base.

The rest of this paper is organized as follows: Section 2 describes the approach and methodology to overcome the challenges of the object relation extraction based on pattern learning. The data set materials and the experiments scenario are elaborated in section 3. Section 4 illustrates the evaluation mechanism that is used to validate the proposed approach. This study uses the accuracy parameter for evaluation purposes. Section 5 contains the explanation of the experiment results and its analysis, while section 6 explains the achievement, the prospects, and the shortcoming of this study. It is also discussed several potential exploration in the future study.

2. APPROACH AND METHODOLOGY

2.1. Objective

The objective of the series of researches is to estimate the potential interaction between entities extracted from unstructured text that is based on its relation. In this study object is synonymous with an entity, since by definition entity is an object or a set of objects in the real world [7]. To achieve this objective, the approach defines a relation between the main object and supporting object. In this relation context, the main objects, for example, are the seller, buyer, drug-name, etc. While the supporting objects are goods or services that were sold by the seller or are bought by the buyer, drug chemistry compound if the main object is drug-name, or children if the main object is father or mother. The global framework of the series of researches can be referred in the previous study [26].

2.2. Pattern Learning of Objects and Its Relation Extraction

In the experiment phase, we do the objects and its relation extraction approach to extract drug name and its component(s) from drug label documents that are distributed in Indonesia Country. The proposed pattern learning approach in this study uses the Indonesian WordNet published by the PAN Localization project [27] as an external knowledge. The Indonesian WordNet is a collection of more than 1.000.000 words in Bahasa Indonesia, which is collected from various sources, such as news agencies, on-line media publishers, Internet blogs, websites and others [27]. Base on the assumption that the names of drugs distributed in Indonesia are unique and are not commonly used in daily term in Bahasa Indonesia, the authors use the Indonesian Wordnet as a guide to determine if certain word identified in the data set is a drug name, a drug component or neither of them. The object extraction framework is illustrated in figure 1. Another input for this framework, in addition to the WordNet, is the initial pattern constructed manually. Initial pattern must contain the relation word. In Bahasa Indonesia, the words that depict the relation between drug-name and drug component are "*kandungan*" (compound) or "*komposisi*" (composition).

Journal of Theoretical and Applied Information Technology
20th April 2015. Vol.74 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

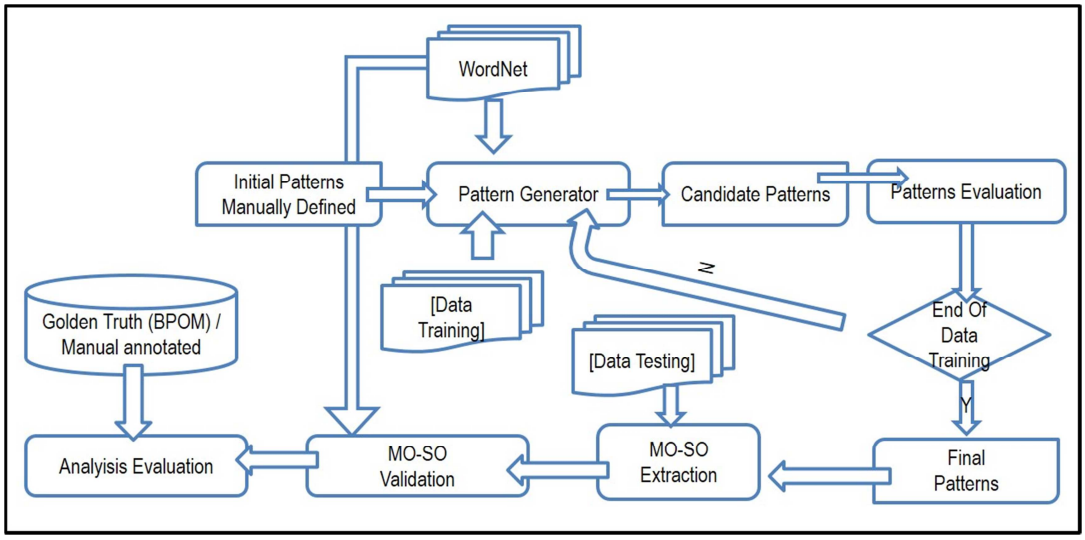

Figure 1. Pattern learning of objects and its relation framework

2.3. Pattern Generation

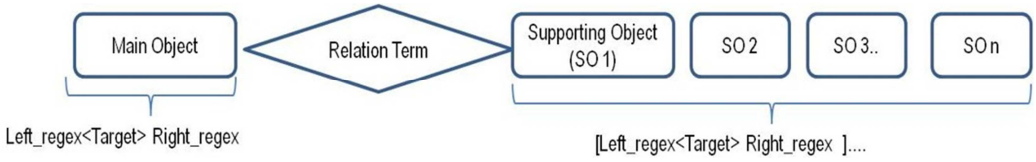

Figure 2. MO-SO relation representation

To extract an object relation (or entity relation), the authors define an object relation as illustrated in figure 2. A relation includes one main object and one or more of its supporting object. In certain document, the relation between the main object and its supporting objects is marked by relation term. The relation term is a certain word, commonly a verb, which describes what kind of relation between the main object and its supporting objects. In the drug-to-drug component relation found in a drug label document, for example, drug is the main object, its chemical components are the supporting objects and "*komposisi*" (composition) is the relation term.

The formal definition of a relation is:

M is a collection of the main object, $M = \{m_1, m_2, ... MP\}$

Si is a collection of supporting object for their main object m_i, $S_i = \{s_{i1}, s_{i2}, ... s_{ic}\}$

A relation r_i is a set of tuples: r_i (m_i, S_i) and R is a set of relation: $R = \{r_1, r_2, ... r_n\}$

Where $m_i \in M$, $s_{ik} \subset S_i$ and r_i is a relation between m_i and s_{ik}.

In drug-name and its component relation, for example, the instance of this relation scheme is:

m_i = *vomilat*; s_i = {*vitamin b6*, *piridoksin hcl*, *folat acid*} and r_i = *"komposisi"*, so the relation can be read as "Drug name *vomilat* contains *vitamin b6*, *piridoksin hcl*, and *folat acid* "

Both of the main objects and supporting objects are identified by a target which are surrounded by a certain regular expression (regex). The regex is performed by left site and right site regex. The pattern is the combination of those regex. In this approach, pattern generation is built by identifying certain words which do not present in the Indonesian WordNet. Once the words are identified, than the other string literals located on its left and right are used as a regular expression of the candidate pattern.

2.4. Pattern Selection

Without any limitation or scoring, the number of candidates' pattern provided in the training phase can be very high. The maximum number of these candidates pattern is as many as the number of

Journal of Theoretical and Applied Information Technology
20th April 2015. Vol.74 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

training document, if each training document contains a unique pattern. If all of the entire patterns that are collected from a large number of documents is used the process, it will consume too many resources (time, processor, or storage). To reduce the size of resources, we select only pattern candidate that fulfills a certain score in testing phase. This pattern score is defined based on the probabilities of left and right regular expression surrounding the target. As explained in the previous study [26], it is performed two pattern scoring scenarios and in this study, we select the best one as described below.

Definition

Each pattern consists of Left Tuple (LT), Relation-Term, and Right Tuple (RT)

Left Tuple Pattern

LT-L = $\{(lt\text{-}l)_1, (lt\text{-}l)_2, \ldots (lt\text{-}l)_n\}$; $(lt\text{-}l)_j$ is j^{th} left regex of the left - tuple

N = Quantity of LT-L, N >= n ; N = n if $(flt\text{-}l)_j$ = 1 for $0<j\leq n$

$(flt\text{-}l)_j$ = frequency of $(lt\text{-}l)_j$ in the training set, then

$$P_j(lt\text{-}l) = (flt\text{-}l)_j / N \,, \quad (1)$$

LT-R = $\{(lt\text{-}r)_1, (lt\text{-}r)_2, \ldots (lt\text{-}r)_m\}$; $(lt\text{-}r)_i$ is k^{th} right regex of the left - tuple

M = Quantity of LT-R, M >= m; M = m if $(flt\text{-}r)_k$ = 1 for $0<k\leq m$

$(flt\text{-}r)_k$ = frequency of $(lt\text{-}r)_k$ in training set, then

$$P_k(lt\text{-}r) = (flt\text{-}r)k / M \,, \quad (2)$$

Left Tuple Pattern Scoring

By treating the left regex and the right regex as independent variables to each other, the number of the left pattern that can be generated are = $J * K$, with J is the number of the left regex of the left tuple and K is the number of the right regex of the left tuple. In the study the scoring scenario is performed based on the condition that there are not all of the right regex are paired with each left regex, but some of the right regexes are belong to certain left regex. This formulation is similar to the conditional probabilities formulation. The formal definition of the left tuple pattern scoring method, $L\text{-}PS_i$, is:

$$L\text{-}PS_i = Pi((lt\text{-}l)_j | (lt\text{-}r)_k) \,, \quad (3)$$

$\exists i,\ i \in \{1,2...n*m\},\ i \leq s \leq n*m;\ \forall j,\ j \in \{1,2...n\}\ ;\ \exists k,\ k \in \{1,2...m\}, k \leq m;$

Right Tuple Pattern Scoring

The Right Tuple Pattern Score, *R-PS*, is derived by the same way with the LPS above. The final formal definition of the right tuple pattern scoring method, $R\text{-}PS_i$, is:

$$R\text{-}PS_i = Pi((rt\text{-}l)_j | (rt\text{-}r)_k) \quad (4)$$

Then the final pattern score, *PS* is:

$$PS_i = Pi((L\text{-}PS)_j | (R\text{-}PS)_k) \quad (5)$$

Algorithm

Based on the object extraction framework and the pattern scoring, the algorithm to generate new pattern by using the training data set is presented in this section. The same skeleton of the algorithm is used both for the Left and Right pattern scoring. The difference is only in the update score mechanism block.

Input (training_set, WordNet, initialPattern)
Output (NewPatternList <Left_Regex, Target, Right_Regex, Prob. of Pattern>)
Algorithm
NewPatternList ←Null
for all documents in training_set **do**
if the document contains initialPattern.relation-term **then**
Get-term-in left of relation-term that is **not** in WordNet
Get Left_Regex, count its frequency
Get Right_Regex, count its frequency for the Left_Regex
for all pairs of Left_Regex and Right_Regex **do**
Perform NewPatternList and Calculate P(R|L)
Reorder on P(R|L) New Pattern in NewPatternList
Output NewPatternList

3. MATERIAL & EXPERIMENT SCENARIO

3.1. Data Set & Pre Processing

The authors collected drugs label documents from various drug producer and regulator web sites as a data set to validate the proposed pattern learning approach. The web sites that are used as the sources of the drug-label documents are http://www.kalbemed.com/, http://www.dechacare.com/, http://infoobatindonesia.com/obat/, http://medicastore.com/, and http://www.pom.go.id/webreg/index.php/home/produk/01. The drug labels are written in Bahasa Indonesia and their common contents are drug name, drug components, indication, contra

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

indication, dosage, and warning. Since the engine grabs almost all of those documents, their format is in htm or html. To filter the real content which contains information regarding drug label, we use html parser provided by http://sourceforge.net/projects/htmlparser/. Total drug label document that is used in this experiment is 15.616 files. The ground truth of the test data is performed manually. The true drug names and drug components are annotated by an expert.

3.2. Experiments Scenario

In this experiment, the data set is split into two parts, one part is as a training set and the other is as a test set by K-fold cross validation with K = 10 respectively. Therefore, there are 10 iterations for each scenario. The generated patterns candidate during training step with data training are sorted in a descending order on its pattern score. Those patterns score is converted to score weight in percentage, so the total score weight of all of the patterns is equal to 1 (100 %). These patterns are applied to the test data set to extract main objects and or its supporting object. Intuitively, it can be seen that the patterns with the higher weight will extract more main objects or its supporting object. The main objectives of this experiment scenario are (1) evaluating the performance of the proposed pattern scoring technique that is represented by the accuracy of the main object and its supporting object extraction; and (2) evaluating the performance of each individual pattern. To evaluate the pattern scoring technique performance, we take the N top of the generated patterns. N is the total percentage of pattern weight, which is in the range from 1 (100%) to 0.5 (50%). In evaluating the performance of the pattern scoring technique, the searching loop to extract the main object and or supporting object will be stopped the first time an object is found. Therefore, it is possible that not all of those patterns are executed and the prioritized pattern to be executed is the pattern that is on the top most of the list. Meanwhile, in evaluating the individual patterns, we use all the patterns generated when trying to find all object-relations. Therefore, all of the patterns will be executed and it is possible to extract certain main object and or its supporting objects more than once.

4. EVALUATION

The performance of all patterns that are based on scoring technique and the individual pattern performance are evaluated base on their accuracy, by using the common criteria in data mining, including precision, recall, and f-score. Let $Cj = \{C_1, C_2, C_3 \ldots Cn\}$ is a set of object extracted by this method against drug-label document set D, and $Ki = \{K_1, K_2, K_3 \ldots K_l\}$ is a set of actual objects in document set D. Adapted from [28], those three criteria computed as follows:

$$\text{Precision}(Ki, Cj) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{|\text{Ki} \cap \text{Cj}|}{|\text{Cj}|} \quad (6)$$

$$\text{Recall}(Ki, Cj) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{|\text{Ki} \cap \text{Cj}|}{|\text{Ki}|} \quad (7)$$

Where $|Ki|$, $|Cj|$, and $|Ki \cap Cj|$ denote the number of object in K, in C, and in both K and C respectively. The computation of F-Score is performed by the formula below:

$$\text{F-Score}\ (Ki, Cj) = \frac{2 * \text{Precision (Ki,Cj)} * \text{Recall (Ki,Cj)}}{\text{Precision (Ki,Cj)} + \text{Recall (Ki,Cj)}}$$

5. RESULTS AND ANALYSIS

The parameter of the accuracy performance evaluation is computed as follows:

$M = \{m_1, m_2 \ldots m_p\}$ *is* a set of main objects extracted from data test by this method, and

$N = \{n_1, n_2 \ldots nq\}$ *is* a set of main object contained in the ground-truth data, then

$$\text{MO} - \text{Precision}(N, M) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{|\text{N} \cap \text{M}|}{|\text{M}|} \quad (9)$$

$$\text{MO} - \text{Recall}(N, M) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{|\text{N} \cap \text{M}|}{|\text{N}|} \quad (10)$$

$$\text{MO} - \text{FScore}\ (N, M) = \frac{2 * \text{Precision (N,M)} * \text{Recall (N,M)}}{\text{Precision (N,M)} + \text{Recall (N M)}}$$

$S_i = \{s_1, s_2 \ldots s_x\}$ *is* a set of supporting object of the main object m_i extracted from test data,

Journal of Theoretical and Applied Information Technology
20th April 2015. Vol.74 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$T_i = \{t_1, t_2 \ldots t_y\}$ *is* a set of main object contained in the ground-truth data, and

$R = \{r_1, r_2 \ldots r_q\}$ is a set of relation contained in the ground-truth data, then

$$\text{SO} - \text{Precision}(T_i, S_i) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{|T_i \cap S_i|}{|S_i|} \quad (12)$$

$$\text{SO} - \text{Recall}(T_i, S_i) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{|T_i \cap S_i|}{|T_i|} \quad (13)$$

$$\text{SOi} - \text{FScore}\ (T_i, S_i) = \frac{2 * \text{Precision}\ (T_i, S_i) * \text{Recall}\ (T_i, S_i)}{\text{Precision}\ (T_i, S_i) + \text{Recall}\ (T_i, S_i)}$$

$$, 1 \leq i \leq q \quad (14)$$

The average *SO-Precision* is

$$AvgSO - Prec = \frac{1}{q}\sum_{i=1}^{q}(SO - Prec)_i \quad (15)$$

The average *SO-Recall* is

$$AvgSO - Rec = \frac{1}{q}\sum_{i=1}^{q}(SO - Recall)_i \quad (16)$$

And the average *SO-FScore* is

$$AvgSO - FScore = \frac{1}{q}\sum_{i=1}^{q}(SO - Fscore)_i \quad (17)$$

5.1. Pattern Scoring Approach Accuracy Performance

5.1.1. First experiment

The first experiment to evaluate the pattern scoring approach is by applying 1 (100%) total score weight of the generated patterns from the training phase to extract the main object and supporting object relations in the testing phase. Using the 100% total score weight, it is possible to use all of the patterns. Because we perform 10-fold cross validation scenario to the test data set, there are 10 iterations in the experiment. By applying of 100% total weight score of a pattern to the test data, the visualization result of all of the ten iterations and the average is illustrated in figure 3 and 4. Figure 3 shows the performance of Pattern Scoring to extract the main object. The maximum performance is achieved in iteration #9 with its F-score = 0.906848426, while the minimum F-score is 0.87913486 in iterations #3, and the average F-score is 0.895959151. The performance of the approach in extracting supporting object is shown in figure 4. The result is quite good since their average F-Score for all the ten iterations is above 0.75. In the supporting object extraction, the best performance is achieved in the first iteration and the worst performance is in the third iteration, with the F-score is 0.815213733, 0.785470285 respectively.

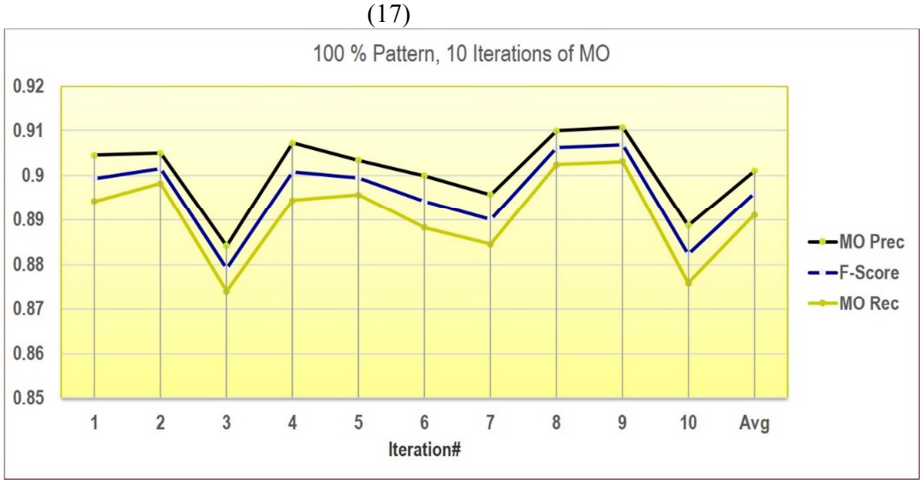

Figure 3. Performance 100% patterns in main objects extraction

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

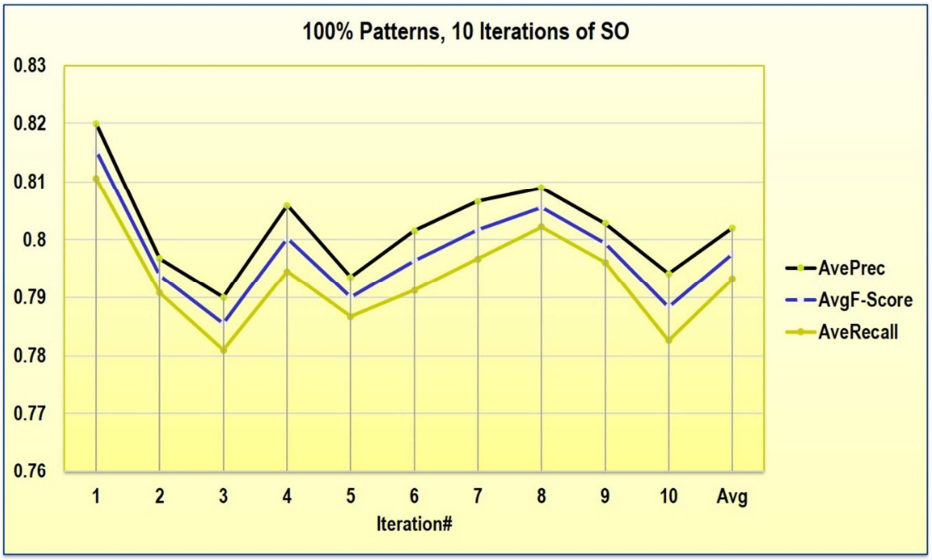

Figure 4. Performance 100% patterns in supporting objects extraction

5.1.2. Second experiment

In the second experiment we execute 5 sub experiments. Each of those sub experiments applied 90%, 80%, 70%, 60%, and 50% of the total score weights of patterns. We did 10 iterations of each these total score weight and compare the results and show the result in figure 5 and 6. The decreasing of total score weight implies that the total number of patterns applied in each sub experiment is also reduced. However, this reduction is not proportional to the total score weight since the patterns which have bigger score weight is located in the top most list, as discussed in the experiments scenario section. Figure 5 shows the performance of each weight score percentage in extracting the main objects. The average precision of 10 iterations is stable from 100% to 50% total weight score, and the value is around 0.9, while their average recall values is decreased. Although the average of recall value decrease as the decreasing of the weight score percentage, but the value is still close to 0.6 for the 50% percentage weight score. The same trend of those performances is shown by the supporting object extraction result in figure 6.

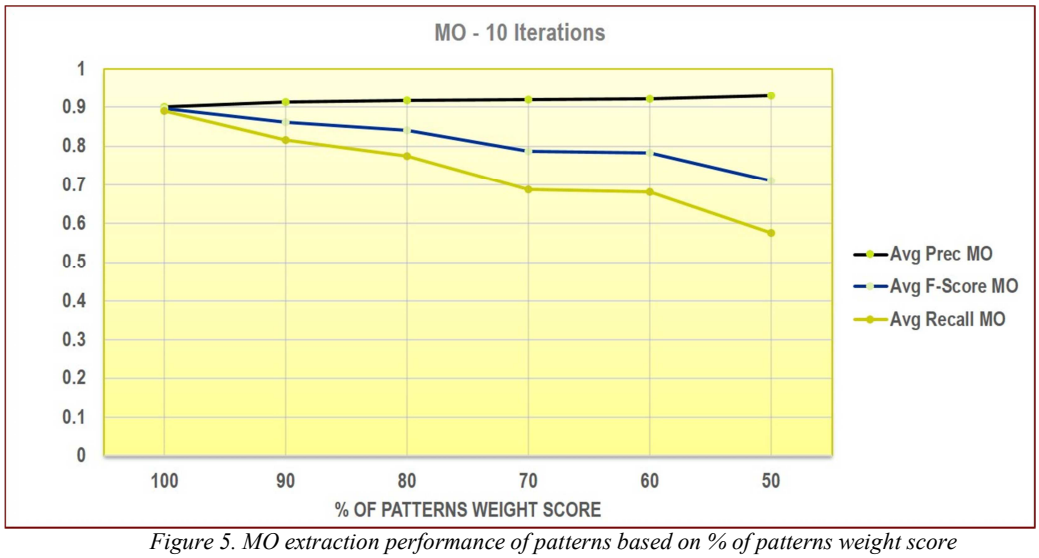

Figure 5. MO extraction performance of patterns based on % of patterns weight score

Journal of Theoretical and Applied Information Technology
20th April 2015. Vol.74 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

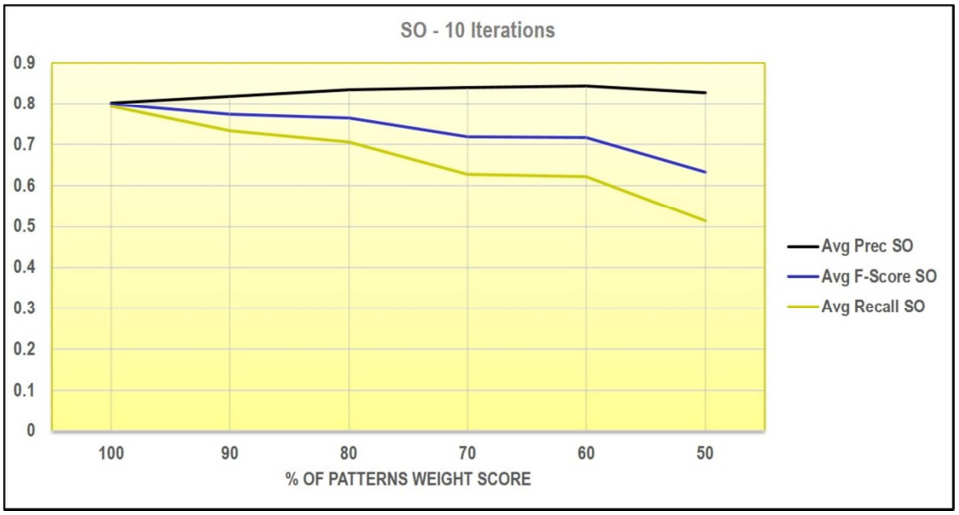

Figure 6. SO extraction performance of patterns based on % of patterns weight score

5.2. Individual Pattern Performance

The evaluation is performed to validate the performance of each generated pattern in the training phase. Since the objective is to evaluate each pattern, then in this experiment each generated pattern is applied to all of test data members. So there are P * D object extraction trials, with P is the total number of pattern and D is the total number of test data members. Because of the generated patterns is quite numerous, then for visualization purposes, the patterns are evenly divided into 4 groups. Each group contains 25% of the total pattern quantity. After those patterns are sorted in descending order based on their weight score, then each of the patterns is identified with a unique number. The grouping mechanism follows the rule as follows: the first 25 % is grouped into Q1, the second 25 % is grouped into Q2, the third 25 % is grouped into Q3, and the last 25 % is grouped into Q4. The parameters used for this evaluation are quantity of object relations extracted and the performance (precision, recall, and f-score). As in the previous experiments, each of the experiment to validate the individual pattern performance is performed in 10 iterations.

5.2.1. The quantity of object extracted

The amount of extracted objects of each patterns group is illustrated as figure 7 and figure 8. These two graphs show the average result of 10 iterations. This result justifies that the pattern selection method and the arrangement of generated patterns base on their weight score is good enough in extracting the object relations. The first 25% group of patterns gives the most significant results, because the first group gives 53% of the total object relation extracted. If we use only the first 50% patterns to extract the relation, it will give around 73% of the total object relations. The last 25% group of patterns gives only 7% of object relations.

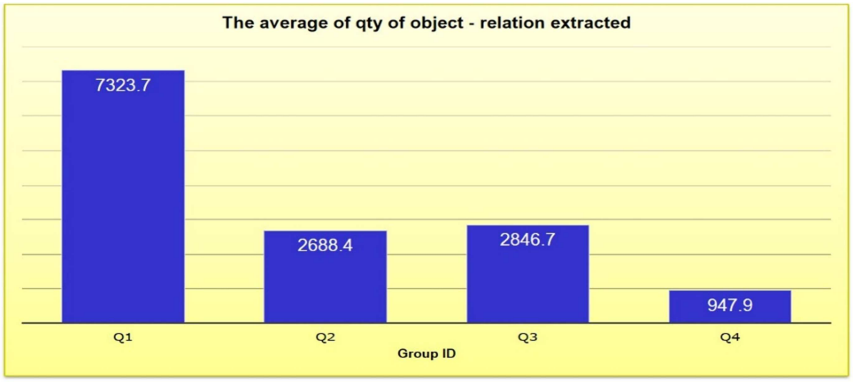

Figure 7. The average quantity of object relation extraction

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

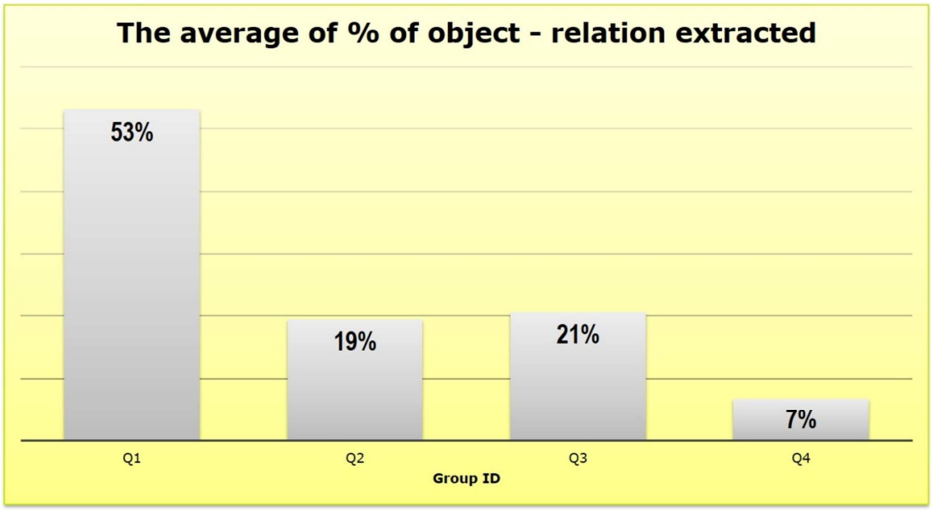

Figure 8. The percentage average of object relation extraction

5.2.2. The accuracy

The second parameter that is used to evaluate the individual pattern performance is their accuracy. The last three figures below show the accuracies of MO and SO extraction results. Figure 9 and 10 show the MO accuracy while figure 11 illustrates the SO accuracy. The way we present the results of MO extraction is different from figure 5. We separate the main object extraction precision with the recall and f-score, because the differences in their value are too high to be shown in the same graph. Their average precisions of the main object extraction are good enough, above 0.75 for Q1 pattern group id, while their recall is very low, so do the f-score. The same trend is also provided by Q1, Q2, and Q3 group as well. The explanation of the high value differences is since this evaluation is applied to each individual pattern which the maximum number of the main object can be extracted by a certain pattern is 501, the minimum number is 0, and the average is 9. Since the number of dataset members are ranging from 1598 to 1699, then the quantity of main objects extracted gives low recall value. But although their recall values are low, their precision, which more than 0.75 for Q1 group ID, is still good enough. Same as the extracted quantity evaluation above, the result of the MO extraction accuracy evaluation also confirms that the pattern selection method, which put the pattern with the highest weight score in the top most list, is feasible to extract the MO-SO relation from the test data set.

The last figure in this paper, figure 11, shows the accuracy of supporting object extraction. For this purpose, each pattern group ID gives relatively similar results. Their F-scores are around 0.24. Even though the F-score of the last pattern group ID, Q4, is the highest when compared to other groups, the difference is not significant since it is only around 0.021. The results also confirms that it is unnecessary to execute all of the resulted pattern to extract the supporting object, because executing a portion of the overall pattern by selecting a certain weight on top of the list will give a better result as is shown in figure 5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

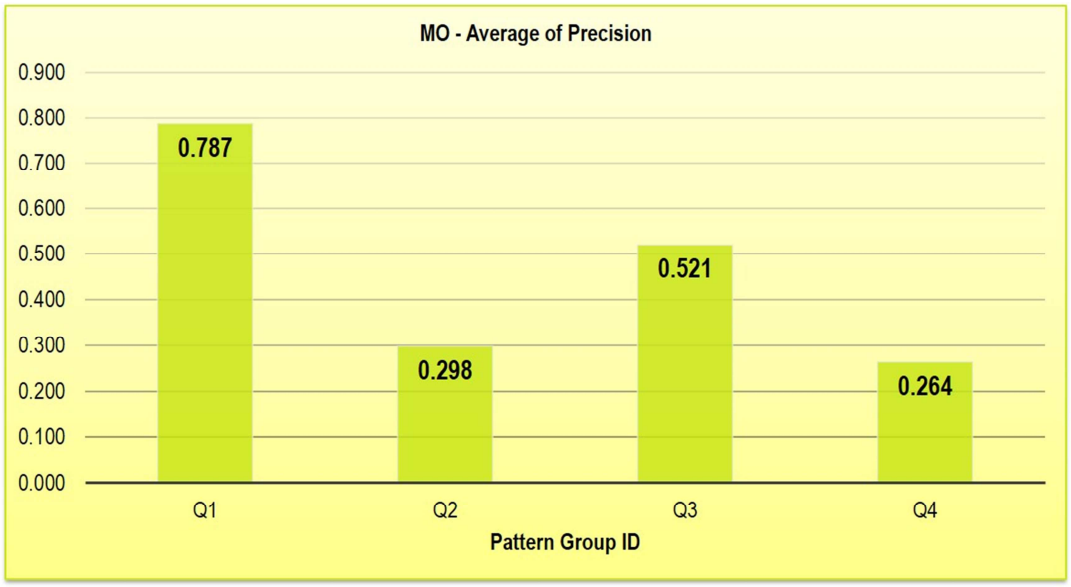

Figure 9. The group ID's precision of the main object extraction

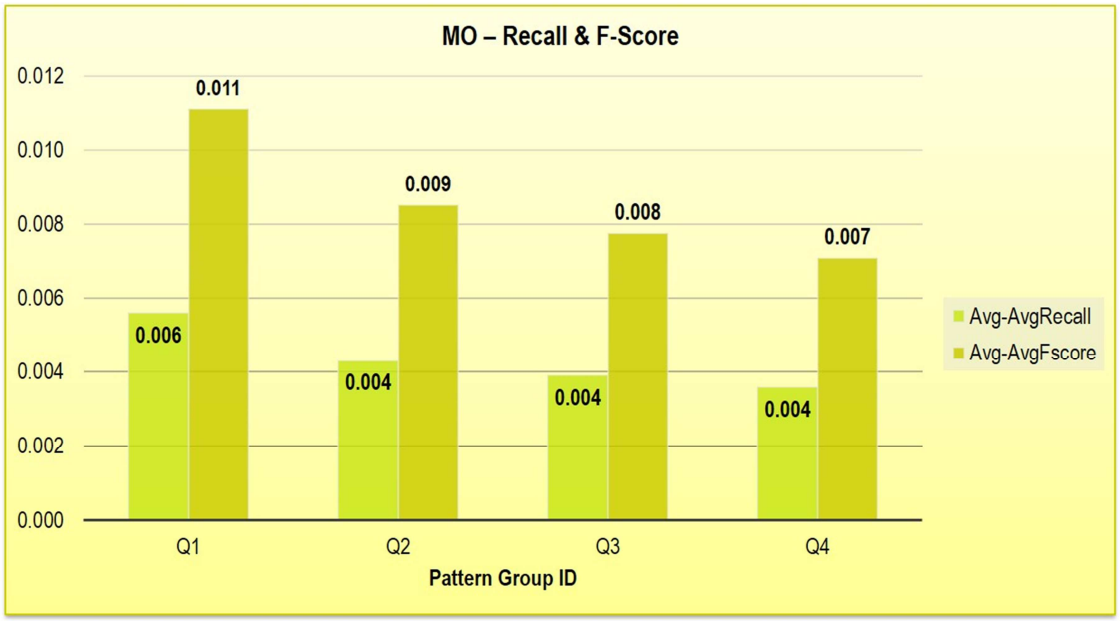

Figure 10. The group ID's f-score and recall of the main object extraction

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

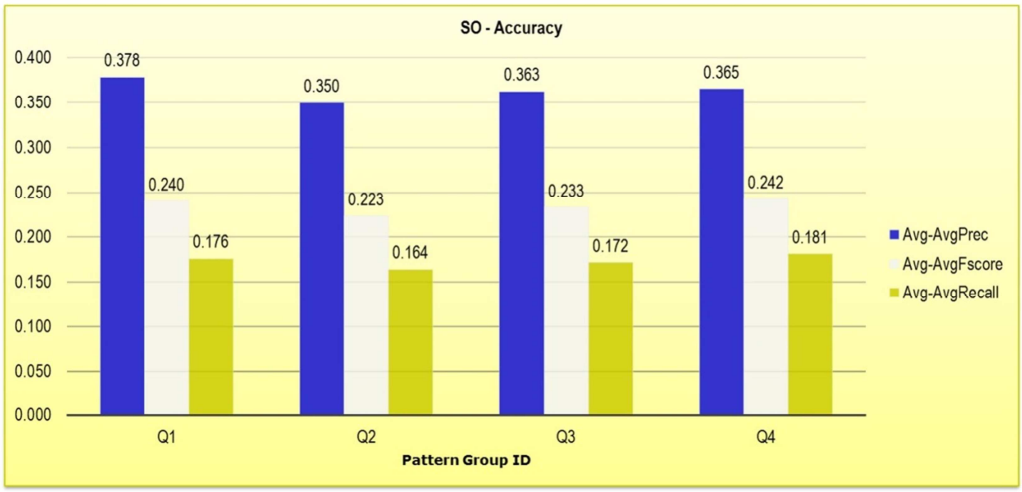

Figure 11. The accuracy of supporting object extraction

6. CONCLUSION & FUTURE WORKS

The study presents the validation of a novel approach to extract object relations which are contained in a semi structured text document. The evaluation of the performances shows that the pattern selection and its arrangement method will be feasible to be explored in the future with some improvements and modifications. The other prospect of the application of this framework is to extract object or relation which is contained in the semi-structured web documents which are generated by the engine. There are plenty of such kind of the document on the web as are shown by almost of all news portals. Although this approach provides several advantages such as no necessity to do some common text preprocessing, the independence to the structure of NLP or to the prior knowledge, but it also has some drawbacks such as dependency on external knowledge or its compatibility limited to semi-structured documents.

In future study, we will explore several potential improvements and modifications such as the usage of test data rather than external knowledge as the guidance of pattern generation, the extension of the test data to other semi-structured documents such as internet based product brochures or news portals. The other prospective improvement methods to extract the main objects which are contained in a certain document is to identify the characteristics of the object including the relative position of the object to the beginning of documents, the frequency of object occurrence in the document, and the word type of the main objects.

REFRENCES:

[1] S.-P. Choi, S. Lee, H. Jung, and S. Song, "An intensive case study on kernel-based relation extraction," *Multimed. Tools Appl.*, vol. 71, no. 2, pp. 741–767, Feb. 2013.

[2] W. Hyperlinks, B. Wei, J. Liu, J. Ma, Q. Zheng, W. Zhang, and B. Feng, "Motif-Based Hyponym Relation Extraction from," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2507–2519, 2014.

[3] Z. Nie, J. Wen, and W. Ma, "Statistical Entity Extraction from Web," *Proc. IEEE*, vol. 100, no. 9, pp. 2675 – 2687.

[4] ACE, "Automatic Content Extraction (ACE) Evaluation," *NIST*, 2014. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/ace/.

[5] N. Bach and S. Badaskar, "A Review of Relation Extraction," in *Literature review for Language and Statistics II*, 2007.

[6] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, and C. Hu, "Mining temporal explicit and implicit semantic relations between entities using web search engines," *Futur. Gener. Comput. Syst.*, vol. 37, pp. 468–477, Jul. 2014.

[7] Y. Chen, "Omni-word Feature and Soft Constraint for Chinese Relation Extraction," in *The 52 Annual Meeting of the Association for Computianal Linguistics*, 2014, pp. 572–581.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] N. Kambhatla, "Combining Lexical , Syntactic , and Semantic Features with Maximum Entropy Models for Extracting Relations," in *The ACL 2004*, pp. 1–4.

[9] L. Qian, G. Zhou, Q. Zhu, and P. Qian, "Relation Extraction Using Convolution Tree Kernel Expanded with Entity Features *," pp. 415–421, 2007.

[10] H. Li, X. Wu, Z. Li, and G. Wu, "A relation extraction method of Chinese named entities based on location and semantic features," *Appl. Intell.*, vol. 38, no. 1, pp. 1–15, May 2012.

[11] T. Wang, Y. Li, and K. Bontcheva, "Automatic Extraction of Hierarchical Relations from Text," in *Proceedings of the 3rd European conference on The Semantic Web: research and applications*, 2006, pp. 215–229.

[12] S. Zhao and R. Grishman, "Extracting Relations with Integrated Information Using Kernel Methods," in *The 43rd Annual Meeting on Associaction for Computanional Linguistics*, 2005, no. June, pp. 419–426.

[13] G. Zhou, L. Qian, and J. Fan, "Tree kernel-based semantic relation extraction with rich syntactic and semantic information," *Inf. Sci. (Ny).*, vol. 180, no. 8, pp. 1313–1325, Apr. 2010.

[14] R. C. Bunescu and R. J. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," in *The Human Language Technology Conference and Conference on Empirical Methods in NLP*, 2005, no. October, pp. 724 – 731.

[15] D. Zelenko and A. Richardella, "Kernel Methods for Relation Extraction," *J. Mach. Learn. Res.*, vol. 3, pp. 1083–1106, 2003.

[16] H. W. Chun, C. H. Jeong, S. K. Song, Y. S. Choi, S. P. Choi, and W. K. Sung, "Composite kernel based relation extraction using predicate-argument structure," *Inf.*, vol. 16, no. 1 B, pp. 657–662, 2013.

[17] M. Collins and N. Duffy, "Convolution Kernels for Natural Language," in *NIPS-2001*, 2001.

[18] H. Tang and J. Ye, "A Survey for Information Extraction Method."

[19] E. Umamaheswari and T. V Geetha, "LEARNING EVENT PATTERNS FROM NEWS TEXT USING BOOTSTRAPPING," in *International Conference on Information System Security And Cognitive Science*, 2013, no. July, pp. 48–54.

[20] E. Agichtein and L. Gravano, "Snowball : Extracting Relations from Large Plain-Text Collections," in *DL '00 Proceedings of the fifth ACM conference on Digital libraries*, 2000, pp. 85–94.

[21] M. Thelen and E. Riloff, "A bootstrapping method for learning semantic lexicons using extraction pattern contexts," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, 2002, vol. 10, no. Emnlp, pp. 214–221.

[22] W. Lin, R. Yangarber, and R. Grishman, "Bootstrapped Learning of Semantic Classes from Positive and Negative Examples," in *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, 2003.

[23] S. Patwardhan and E. Riloff, "Learning Domain-Specific Information Extraction Patterns from the Web," in *Proceedings of the Workshop on Information Extraction Beyond The Document*, 2006, no. July, pp. 66–73.

[24] T. Liu and T. Strzalkowski, "Bootstrapping Events and Relations from Text Polish Academy of Sciences," in *The 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, no. 1996, pp. 296–305.

[25] R. Huang and E. Riloff, "Multi-faceted Event Recognition with Bootstrapped Dictionaries," in *Proceedings of NAACL-HLT 2013*, 2013, no. June, pp. 41–51.

[26] M. Sadikin and I. Wasito, "Toward Object Interaction Mining By Starting With Object Extraction Based on Pattern Learning Method," in *2014 Asia-Pacific Materials Science and Information Technology Conference (APMSIT 2014)*, 2014.

[27] R. P. Secretariat, "PAN Localization Project, Indonesia country component," 2010.

[28] G. G. Dagher and B. C. M. Fung, "Subject-based Semantic Document Clustering for Digital Forensic Investigations," *J. Data Knowlege Eng.*, vol. 86, no. October 2013, 2013.