Towards Effective Entity Extraction of Scientific Documents using Discriminative Linguistic Features

Sangwon Hwang[1], Jang-Eui Hong[2], Young-Kwang Nam[1*]
[1]Department of Computer & Telecommunications Engineering, Yonsei University
Wonju, South Korea
[e-mail: arsenal@yonsei.ac.kr, yknam@yonsei.ac.kr]
[2]Department of Computer Science, Chungbuk National University
Cheongju, South Korea
[e-mail: jehong@chungbuk.ac.kr]
*Corresponding author: Youngkwang Nam

Received October 22, 2018; accepted January 17, 2019; published March 31 2019

Abstract

Named entity recognition (NER) is an important technique for improving the performance of data mining and big data analytics. In previous studies, NER systems have been employed to identify named-entities using statistical methods based on prior information or linguistic features; however, such methods are limited in that they are unable to recognize unregistered or unlearned objects. In this paper, a method is proposed to extract objects, such as technologies, theories, or person names, by analyzing the collocation relationship between certain words that simultaneously appear around specific words in the abstracts of academic journals. The method is executed as follows. First, the data is preprocessed using data cleaning and sentence detection to separate the text into single sentences. Then, part-of-speech (POS) tagging is applied to the individual sentences. After this, the appearance and collocation information of the other POS tags is analyzed, excluding the entity candidates, such as nouns. Finally, an entity recognition model is created based on analyzing and classifying the information in the sentences.

Keywords: Named entity recognition, entity extraction, data mining, data cleaning, sentence segmentation, information extraction

A preliminary version of this paper was presented at APIC-IST 2018, and was selected by the conference review process. This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2014M3C4A7030503) and (in part) the Yonsei University Research Fund of 2014..

http://doi.org/10.3837/tiis.2019.03.030 **ISSN : 1976-7277**

1. Introduction

In the current era of big data, numerous types of information are available in a variety of digital forms. Such information can be structured or unstructured and often needs to be extracted and processed during various preprocessing tasks of natural language processing (NLP). While developing technologies for processing big data, the extraction of meaningful information from large-scale data can be a challenging issue. Several applications of information extraction (IE) and NLP require certain preprocessing tools for analyzing textual components such as lexical and morphological elements.

NER [1] is an important technique in the field of IE [2] for identifying named entities (NE) that appear in various forms in a text document. An NE can represent a single piece of information, such as the name of a person, place, technology, or product, and can be used to identify and classify the subject of the document. In the case of applications involving information retrieval (IR) and question answering (QA), the quality of the search results depends on the results of NER, and the associated search system thereby has greater reliability. Furthermore, among the emerging technologies announced by Gartner in July 2013, the techniques of natural language question answering (NLQA) and complex event processing (CEP) are also based on NER [3].

Many methods have been developed to identify NEs in documents [4], such as feature based clustering [5] and methods to improve the recognition performance of Korean named-entity recognition by clustering word vectors via the k-means algorithm [6,7]. The word-vector clustering technique has also been applied to entity recognition problems in the medical domain [8]. The authors in [9] reviewed three biomedical NER models and two machine learning methods, namely, the hidden Markov model (HMM) and conditional random fields (CRFs)). The authors in [10] proposed a method of extracting NEs via a probabilistic model based on the maximum entropy (ME). However, a limitation is that these methods use preprocessing to eliminate the stop-list and select nominal words that have undergone morpheme analysis as features. In contrast, the method proposed in this paper extracts entities by analyzing the frequency of co-occurrence of the stop-list, which was excluded from previous studies. A functional keyword (FK) consists of a combination of stop-words that appear together with the entities in various forms in the context, and can be used during contextual analysis to extract entities from the text and identify the attributes of the categories (e.g., technology, theory, person) to which the entity belongs. Moreover, the method can be used to identify relationships between entities in various systems, such as question answering systems. The proposed method was validated using 4,200,667 article abstracts obtained through the Elsevier developer portal (EDP).

In section 2, we review the previous work briefly. We describe our entire system in section 3 and report the implementation and experimental result in section 4. Section 4 reports the implementation and experimental result, Finally, we describe result and future works in section 5.

2. Related Work

IE techniques are used to extract structured information from unstructured and semi-structured machine-readable documents. This information primarily consists of noun-type entities identified via NLP algorithms in text based human language. In [11], the authors defined an NE as "a proper noun that serves as a name for something or someone." NEs were defined as "unique identifiers of entities" at the 6th MUC Conference in 1996 [12]. In [13], the authors classified NEs based on their grammar, rigid designation, unique identification, and domain of application. Similarly, NEs in the current study are defined according to their application and domain.

NER techniques are used in a variety of applications, such as semantic annotation, to improve information extraction and interoperability [14]. For example, if the ontology of the city concept called Seoul is linked to the country concept of Korea, the ambiguity of the word Seoul in the text can be avoided. Identifying such relationships in large-scale collections via automated NER techniques is one way to reduce the burden of annotating new documents [15].

The NER technique is also used in QA systems as a method of locating specific answers to queries. In fact, approximately 80% of the queries analyzed at the Eighth Text Retrieval Conference (TREC-8) sought information regarding who, where, and when, and the responses referred to entities in the form of people, organizations, and regions [16].

Understanding the ontology in a particular domain is important for applications that must interoperate with the semantic web. Thus, it is essential that it be possible to quickly generate a simple ontology without human intervention [17]. When such systems are being integrated, NER techniques can be incorporated to automatically identify and implement many of these functions. One example of a tool in which many of these techniques have been implemented is the KnowItAll [18] tool.

In recent years, NER has also been used in the social web. When people are asked for their opinions or make decisions, final decisions are often made based on information retrieval [19]. One preprocessing technique in opinion mining is the recognition of the sentiment of responses to determine if they perceived as positive or negative. The OPINE system was developed to analyze interrelated opinions and extract the corresponding sentiment toward the product [20].

NER also plays an important role in machine translation where it is used to automatically translate source texts or speech to a target language [21,22]. The misidentification of NEs not only affects the lexical recognition accuracy, but also the context as a whole, which then affects the translation quality. Automatic NER techniques can be used to improve the quality of machine translation systems.

Automatic text summarization refers to the use of NE equipped software tools to shorten text documents by creating summaries of the key points in the original document. In such applications, the quality and performance of the NER system significantly affects the quality of the summary [23,24]. Moreover, if weighted scores are employed when

extracting the NEs, the number of NE iterations required to summarize documents can be reduced [25].

Text clustering is used to group text with common properties into a cluster or group, and is primarily used in the domain of knowledge search and data mining. This technique includes keyword extraction and NER, and is used to identify the type of person, region, institution, and words contained in the text. In [26], a combination of NE and keyword based techniques was employed to improve the quality of clustering, and in [27], NER was used to improve the performance of suffix tree clustering for new search results.

In the field of information retrieval (IR), NER has been applied to determine the intent of the user that is implied in the query. An appropriate keyword and entity search can enable high quality search results, which can be used to quickly locate information in large-scale databases [28]. Since IR systems do not analyze the information to be searched, the addition of robust NER techniques simplifies and improves the accuracy of IR systems.

NER has also been utilized in the medical field to assess interactions between drugs [29,30], the action and side effects of drugs [31,32], diagnostic classifications [33], and in the search and classification of biomedical entities [34].

From a technical perspective, NER approaches can be categorized as rule-based, learning-based, and hybrid approaches. In the first rule-based system, NEs were classified based on handwritten syntactic-lexical rules [4]. This proved to be efficient as it employed the properties of knowledge related to the language [35] and utilized domain specific features to improve accuracy. However, this approach is costly and domain-dependent, and cannot be applied in other domains. In addition, this implementation required the programming skills and intervention of domain experts [36].

Learning-based approaches use learning models and various algorithms to automatically extract NEs in a supervised, semi-supervised, or unsupervised manner. Supervised-learning first requires a domain expert to label the data in the training dataset and to select appropriate algorithms to use during learning. Once this is complete, the system can be used to analyze new data. Examples of ML-based approaches are hidden Markov models (HMMs) [37,38], support vector machine (SVM) based systems [39], conditional random field (CRF) based systems [38,40,41,42,43,44,45], maximum entropy Markov models (MEMMs) [46,10,47], logistic expression based systems [48], and some classifier ensemble techniques [49]. A disadvantage of these methods is that they require a large amount of training data and have a lengthy run time.

In semi-supervised learning, a small amount of data is repeatedly relearned while unlabeled data is tagged to compensate for this shortcoming. Unsupervised learning techniques do not rely on training or answers when forming result groups, although hidden features or structures can be found in the results. The NEs in such groups have a probabilistically high correlation. The authors in [50] proposed a method of extracting NEs based on the vector distance computed using the Word2vec model in a particular document. In [51], the authors proposed a model for improving the quality of NER systems by using latent semantics.

Hybrid approaches combine the advantages of both rule-based and learning-based techniques. For example, the authors in [52] combined a CRF and a post-processing algorithm to extract biomedical entities, and the authors in [53] implemented a Turkish NER system that combined lexical resources, patterns, and role-based learning. Jason et al. developed a system to automatically extract features based on words and characters by combining bidirectional long-short term memory (LSTM) and a convolutional neural network (CNN) [54].

State-of-the-art approaches eliminate the stop-list, which is considered to be unnecessary when analyzing various objects, such as "prepositions," during the preprocessing stage. The key to such tasks is tagging or labeling objects based on the respective noun types; however, this requires the manual intervention of the developer or system manager before unsupervised learning can be applied, which is costly and time consuming.

In the current study, a method of analysis using discriminative linguistic features is proposed to overcome the problems in the above listed approaches. The proposed approach can be used to extract entities by combining the stop-lists that were removed in previous studies and identifying the relationships between the entities. Combinations of words, such as prepositions and adverbs, that appear frequently and simultaneously are often key to extracting idioms and phrases. The information that appears in a sentence can be identified based on the extracted entities and their properties.

This combination is defined as an FK. In a similar study [55], the authors presented a method for structuring biomedical abstracts by analyzing the characteristics of the language. With the model presented in [55], abstracts from MEDLINE in an unstructured format were restructured into an introduction, methods, results, and discussion (IMRAD) [56,57] format. The analysis model was created by combining model and tense verbs without noun types. That study focused on analyzing the regional characteristics of the IMRAD sections in the resulting structured document and created models that could be used to structure unstructured abstracts.

In contrast, in the current study, FKs and information in the IMRAD format are used to analyze the entities within each sentence rather than to structure the entire document. This information can also be used when analyzing and classifying the relationships between extracted entities. The proposed method is described in detail in the next section.

3. Methodology

IE refers to the extraction of structured information from unstructured and semi-structured machine-readable documents via natural language processing (NLP) techniques. The majority of the extracted information is in the form of noun-type entities. In [11], the authors defined an NE as "a proper noun that serves as a name for something or someone." NEs were referred to as "unique identifiers of entities" at the 6th MUC conference in 1996. In [13], the authors defined NEs by classifying each based on their grammatical category, rigid designation, unique identification, and domain of application. With that approach in

mind, in this study, NEs are defined according to the specific application and domain.

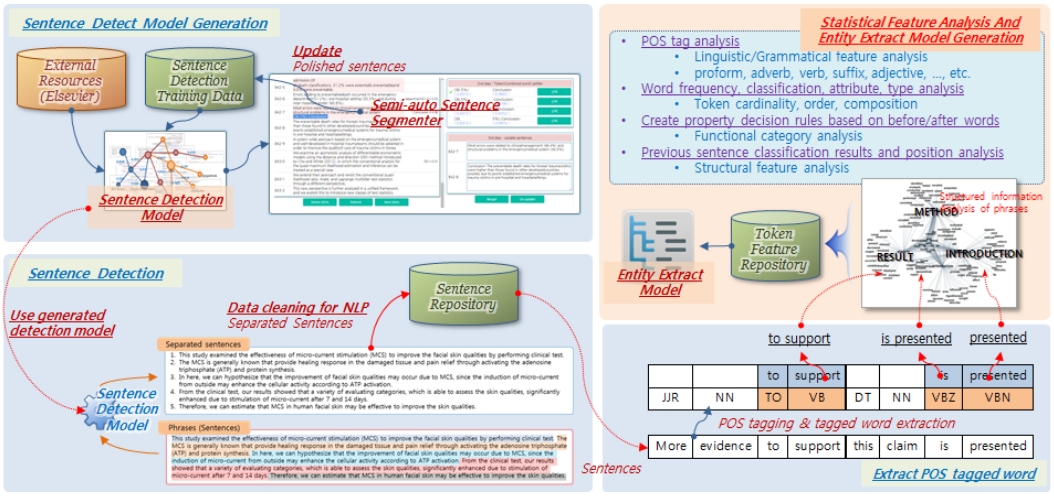


Fig. 1. System overview

An overview and flowchart of the proposed system is shown in **Fig. 1**. As shown, the system incorporates two analytical models: a sentence detection model, which uses a sentence detector (SD) to divide the text into unit sentences, and an entity extraction model, which uses an entity extractor to extract entities from objects. The individual functional modules that make up the complete system are shown in **Fig. 2**. These can be divided into search functions and training functions. The search part collects document data, such as journal papers or articles, while the training part creates the model used to split paragraphs into individual sentences and to extract entities using the data collected by the search part.

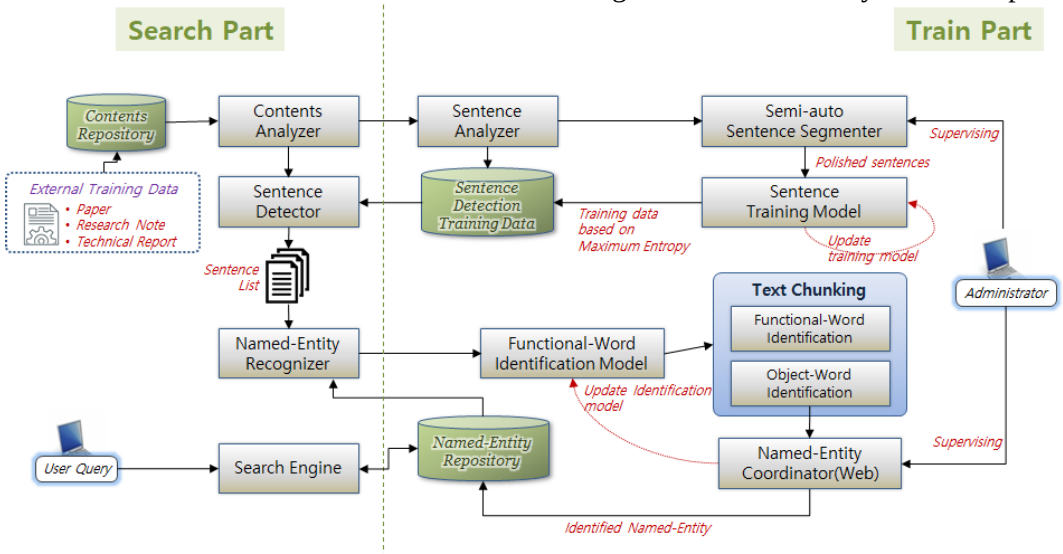


Fig. 2. System architecture based on functional modules

The system can be used to process the abstracts of journal papers, research notes, technical reports, and theses. In such documents, abstracts summarize the key elements of the main document so that readers can quickly gain an understanding of the material covered [58] in order to decide if further attention is warranted. With this in mind, abstracts can be regarded as a set of high-value sentences describing the background, theory, purpose, method, and results of the research described in the main document. The noun-type entities included in each sentence include techniques, algorithms, names, and so on. The focus of the current study was to identify the FKs that appear around entities to define or otherwise support the relationships between entities.

The proposed algorithm was applied to the collection of documents in order to detect the constituent sentences, and the results were then used as the training data for the sentence detection model. The administrator rebuilt the model by correcting the errors in the sentences extracted by the semi-automatic SD. NLP was then applied to the detected sentences in order to generate an entity analysis model via statistical analysis. The following subsections describe the details of the function modules.

3.1 Sentence Analysis

In data mining applications, data cleansing is essential to improve the quality of the final result. For NER in particular, the quality of the results depends on the quality of the raw data. The documents analyzed in this study consist of the abstracts of journal articles, each of which is made up of multiple sentences. During sentence analysis, the preprocessor separates each paragraph into single sentences. This is depicted in **Fig. 3**.

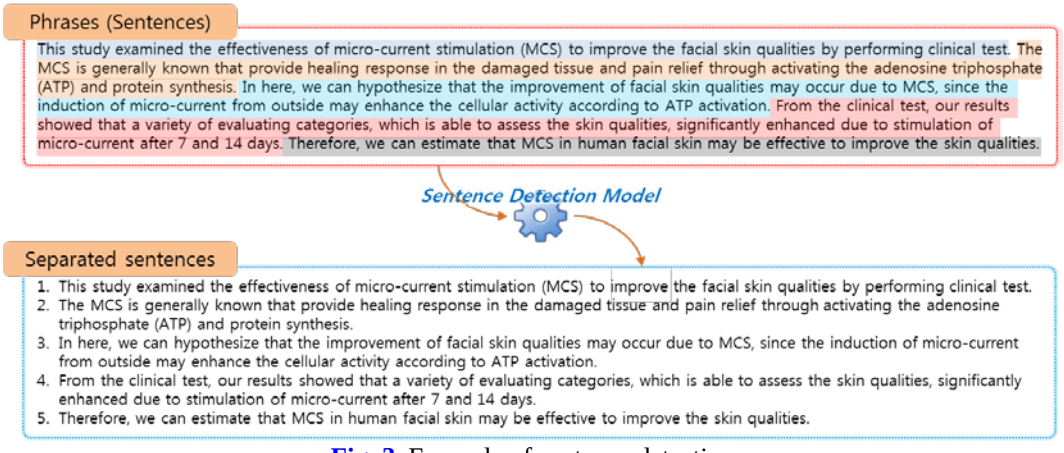


Fig. 3. Example of sentence detection

While periods are generally used to terminate sentences, they are also used in other ways, such as in abbreviations (“T. J. Harper.,” “I.B.M.,” etc.), and numeric values (“0.5%,” “3.14159,” etc.). If sentences containing such words are tokenized based only on the

locations of the periods, the resulting sentences may appear garbled or otherwise incomplete. To overcome this issue, the SD in the proposed method employed the sentence learning model provided by the Apache Software Foundation OpenNLP library to properly separate the sentences via a pre/post-processor.

Several techniques are available for separating sentences. The first is the sentence splitting function in CoreNLP, which separates sentences based on the PennTreeBank (PTB) dataset. The second is the SD in OpenNLP that uses a supervised learning technique based on the maximum entropy (ME) [57]. Although many machine learning algorithms, such as Naïve Bayes [58], can be used for supervised learning, in the current study, ME was used due to its slightly better performance in a single domain.

To improve the performance of sentence detection, the model was calibrated by repetition using a semi-automated sentence segmentation (SSS) system (see **Fig. 4**) developed for this purpose.

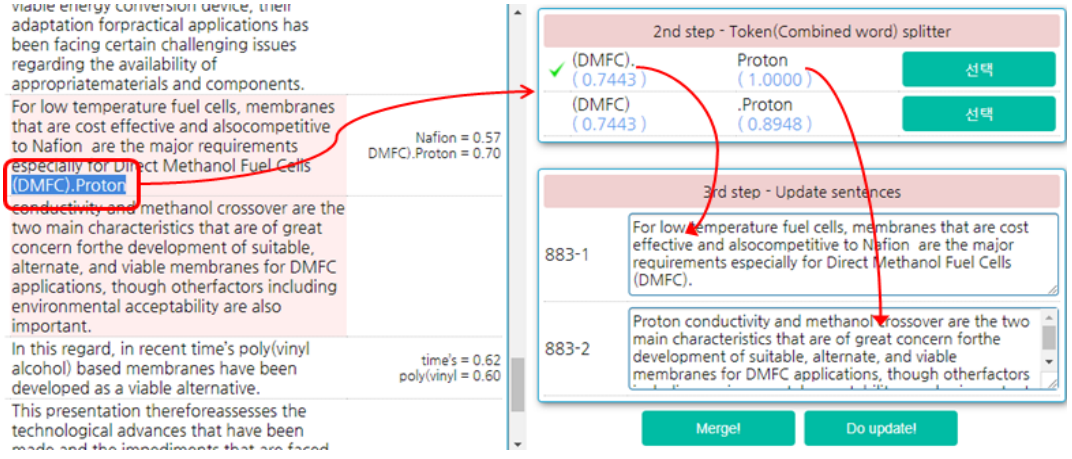


Fig. 4. Example of semi-automated sentence segmentation system

First, automatic sentence detection was performed on the collected documents, after which the administrator verified and calibrated the results using the SSS system. During operation, the left-hand side of the screen displayed the sentence detection results and verified that each sentence was properly separated using the maximum-entropy-based token probability model.

The token probabilistic model displays the probabilistic values on the screen after determining whether or not the word in the sentence was used appropriately. The model was generated via ME-based learning using various words as training data. When the token probabilities were calculated for the following two sentences, the results are shown in **Table 1**.

- "Winter is coming.It is the motto of House Stark."
- "Winter is coming. It is the motto of House Stark."

Table 1. Token probabilities for the two sentences

Token	Tag	Probability	Token	Tag	Probability
Winter	NNP	0.631323119	Winter	NNP	0.547055175
is	VBZ	0.971406981	is	VBZ	0.988052539
coming.It	**NN**	**0.359071061**	**coming**	**VBG**	**0.954201859**
is	VBZ	0.888892686	**.**	**.**	**0.367869324**
the	DT	0.986599644	**It**	**PRP**	**0.965098893**
motto	NN	0.944683589	is	VBZ	0.994607214
of	IN	0.958671464	the	DT	0.989590443
House	NNP	0.984494406	motto	NN	0.944683589
Stark	NNP	0.991990207	of	IN	0.958671464
.	.	0.974601623	House	NNP	0.984494406
			Stark	NNP	0.991990207
			.	.	0.974601623

These sentences are similar, except that the first includes "coming.It" in the middle while the second includes "coming. It" with an additional space in the middle. The left side of Table 1 lists the probabilities of each token in the first example sentence, most of which were above average. However, the probability of “coming.It” is below average, which indicates that the token is probabilistically invalid. In contrast, it can be seen that in the right side of Table 1, the probability value of “coming” and “It” are sufficiently high.

The SSS system calculates the probabilities for each token in the resulting sentences and displays the probabilities of tokens with values below the threshold (0.8). These tokens were identified based on several special characters (e.g. ‘.’, ‘;’) when determining the probability values for each separated word. If two split words exceed the threshold (0.85) required to exist as an independent word, the previous token is appended to the end of the previous sentence and the next token is added to the beginning of the following sentence in order to complete the two sentences. Tokens that could not be automatically separated can be manually separated by the administrator in the management system. The sentence detection model was constructed by learning 31,748,498 sentences that were extracted from 4,200,667 abstracts.

3.2 POS tagging

The POS tagging module tags each unit sentence and extracts the parts of speech that can be used as FKs (e.g. VBs (verbs, base form), VBDs (verbs, past tense), VBZs (verbs, 3rd person singular present), RBs (adverbs)), and entities (e.g., NNS (nouns, plural), NNPs (proper nouns, plural), JJRs (adjectives, comparative), JJSs (adjectives, superlative)). If a determiner (DT) appears before an NN, the tag is combined with the NN. The Stanford CoreNLP POS tagger (POSTagger) was used in this study as it extracts POS tags with high accuracy.

3.3 Entity Analysis

During entity analysis, various tasks, such as linguistic and grammatical feature analysis, are performed based on the POS information to generate statistical data, such as token cardinality, order, co-occurrence, collocation, and composition. Then, the dependency information among the tokens is used to create property determination rules based on the pre-/post- words. After this, the position of the sentence and the structural features based on the before/after sentences are analyzed. Then, FKs are generated by considering word groupings, such as "be-able-to" and "in-order-to." This is necessary as the meanings of "be," "able," "to," "be-able-to" are different, and therefore must be separately processed. The system first measures the frequency of appearance of the word combination "be-able-to," and then determines the frequency of occurrence of "be," "able" and "to." When the frequency analysis is complete, a list of related collocated words and their frequencies is generated.

The collocation relationship between words is quantified using the likelihood ratio, which is defined as follows. Let w be a word, p be a probability, and H be a hypothesis. Then,

H_1. The probability p_1 of occurrence of w_2 when w_1 also occurs is equal to the probability of w_2 occurring when w_1 does not occur.
H_2. The probability p_1 of occurrence of w_2 when w_1 occurs is not equal to the probability of occurrence of w_2 when w_1 does not occur.

Hypothesis H_1 applies when the occurrence of w_1 and w_2 are independent, while H_2 applies when the occurrence of w_1 and w_2 are mutually dependent. The likelihood ratio of a hypothesis can be calculated as follows:

$$Likelihood\ Ratio\ \lambda = \frac{Likelihood\ (H_1)}{Likelihood\ (H_2)} \qquad (1)$$

Table 2 lists the likelihood-ratio hypotheses of H_1 and H_2, where N is the length of the entire abstract, c_1 and c_2 denote the number of occurrences of w_1 and w_2, respectively, and c_{12} indicates the number of co-occurrences of the two words.

Table 2. Likelihood-ratio hypotheses [59]

	H_1	H_2
p_1	$\frac{C_2}{N}$	$\frac{C_{12}}{C_1}$
p_2	$\frac{C_2}{N}$	$\frac{C_2 - C_{12}}{N - C_1}$

The likelihood λ can be transformed into a chi-square distribution by applying the equation –2logλ [60]. Then, two words that have degrees of freedom equal to one and a critical value (a = 0.005) greater than 7.88 are considered to be in a collocation relationship. In this study, this is defined as a FK, which can be used during IMRAD [56,61] classification.

The generated FK determines the category based on the IMRAD classification method. For example, in the "introduction" category, several expressions of "to" infinitives, such as "to determine," are used to explain the purpose of the study. The FKs are used to create creates a model for extracting the properties of entities (e.g. techniques, products, research methods, etc.) that are represented by the words before/after the functional words in a certain classification. As an example, **Fig. 5** shows the results of an entity analysis of the following sentence: "Naive Bayes algorithm and Support Vector Machine (SVM) are used for fine-grained emotions classification of tweets."

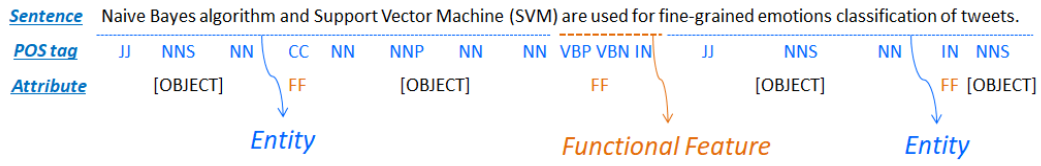


Fig. 5. Example of entity analysis stage

The words of type VBP (Verb, non 3rd person singular present), VBN (Verb, past participle), or IN (Preposition or subordinating conjunction) can be FKs, while the others can be entities. In **Fig. 5**, the term "are-used-for" implies that the entity that appears before it can be a tool or method for the entity that appears after it. In this way, FKs are used in the classification of the relationships between entities during extraction.

Based on the techniques of identifying collocation relationships and tag-based analyses, the cluster analyzer creates an FK list and FK clusters using the following procedure. First, the system extracts the tags that include the FKs along with matching words. Then, consecutive tags are concatenated into one FK. These are extracted as shown in **Fig. 5**. This procedure is repeated for all of the sentences to compute the appearance frequency of the tag pattern. An example of word groups containing specific tag patterns is shown in **Table 3**. In the table, "ETT" indicates that a word or word-set that can be an entity.

Table 3. Sample of n-gram tag patterns

n-gram	Tag pattern (TP)	Freq. of TP	Sample sentence
1	**IN** ETT	41,947,810	**By** laser **for** breast cancer recurrence **of** carbonic anhydrase
	CC DT ETT	2,355,641	**and a** Western blot method **and the** intensification **and a** perfect parabolic surface
	ETT **VBD VBN IN DT** ETT	896,634	Higher pressure **was obtained at the** PP location DSW1500 drinking waters **were formulated via a** combination luminescence **was analyzed within the** frame work
2	ETT **CC** ETT **IN DT** ETT	1,917,240	hydrogen **and** oxygen **from the** air design assistance **and** process support **during the** design phase kinematics **and** femoral liftoff **on the** wear
	ETT **VBD DT** ETT **IN** ETT	174,254	work **investigated the** impact **of** lactose crystallization third class **modeled a** papercraft **with** instruction Caffeine withdrawal **caused a** decrease **in** Adora1 mRNA level
	ETT **VBG DT** ETT **IN DT** ETT	156,216	Interpretations **regarding the** meaning **of the** grooves tuples **using the** notions **of the** first homology group value **limiting the** laminar regime **for the** incoming flow
3	ETT **IN DT** ETT **IN** ETT **IN** ETT	292,602	useful **in the** legislative tools **for** implementation **of** energy conservation measures turn **of this** century **as** labels **for** new research approaches particle shape **as an** important tool **for** passive targeting **of** nanocarriers
	ETT **VBZ** ETT **IN** ETT **CC** ETT	29,028	IMT-MBG system **is** effective **for** bone tissue regeneration **and** bone cancer treatment physical exercise **is** beneficial **for** motor **and** non-motor symptoms free drug concentration **is** similar **in** plasma **and** tissue
	ETT **DT** ETT **CC** ETT **IN** ETT	25,322	engineering **the** variable **and** constant regions **of** antibody fragments modeling **the** knowledge **and** deduction methods **of** experts functional unit **the** rotor blades **and** packages ready **for** installation

Of the results in **Table 3**, since “IN ETT” and “CC DT ETT” are prepositions and conjunctions that include “by~” and “and~” they do not make much sense on their own. Thus, the tag patterns that cannot be utilized or are otherwise meaningless are removed at the preprocessing stage. The above results are an example of a 1-gram FK. The system performs an n-gram analysis, where the n of the n-gram FK indicates the number of FKs that exist in the FK pattern. In this study, n <= 5.

When the n-gram pattern extraction is completed, clustering is performed based on the FK for the high-frequency pattern. Frequent patterns are often used in sentences, and there are various detailed classifications for those patterns. For example, because the pattern "WDT" might have different classifications of entities that contain "that" and "which," it is necessary to create each analytical model for "an apparatus which includes pumping system" and "method that create or remove naming tool." Therefore, for each sentence pattern, create a group using the k-means++ and create an analysis model based on each group. This improves the disadvantage of the k-means algorithm, which depends on the selection of initial values [62]. The following table shows examples of analytical models.

Table 4. Example of analytical model

Tag pattern	Constraints	Results
ETT1 VBD VBN IN ETT2	ETT1 ∈ {“analysis”, “study”, “approach”, “test”, …} VBD ∈ {“was”, “were”} VBN ∈ {“applied”, “assessed”, “conducted”, “performed”, “treated”, “used”, ...}	ETT1={method} ETT2={property| technique}
	VBD ∈ {“was”, “were”} VBN ∈ {“achieved”, “associated”, , …}	ETT1={entity} ETT2={property| value}

Regarding the meaning of tag pattern in **Table 4**, if FK contains a "VBD VBN IN" type tag and exists before "ETT2," "ETT1" can be "entity" or "method" and "ETT2" can be the "value", “technique” or "property." In the following sentence, “skin cancers were treated with medical oncology wards,” “skin cancers” can be an object representing “entity” or “entity candidate” and “two separate study areas” can be “technique” or “property.”

3.4 NER Training Model Generation

The entity identification model can be used to analyze sentences and extract the various entities in a sentence. These include named entities, which have recognized names, and entities that have not been identified as entities but can be identified as entities nonetheless. In this study, the NER module was developed using the Apache OpenNLP library. The NER module manages the various entities that have been classified as NEs that are extracted with the entity identification model. If the extracted entity in the previous step is unclassified, it can be classified as an NE through its learning system. For example, in the following sentence, the classification system can extract entities in the form of a technology, algorithm, or certain tool type depending on the “Methods” classification.

<START:algorithm>Naive Bayes algorithm and Support Vector Machine (SVM)<END> are used for fine-grained emotions classification of tweets.

As shown in the above sentence, the model generation system adds <START>/<END> annotations to the front and back of the extracted entity to generate training data for use during NE learning. The learning system performs NE learning using the training data generated in the previous stage to construct the NER model. This model enables the NER system to extract NEs or significant entities from the abstracts in various scientific articles. In this study, a semi-automated entity-tagging technique was used to tag entities and generate training data.

4. Extracting Model Generation

The system was developed using Java 1.8, and the OpenNLP and CoreNLP libraries were used for the NLP functions. A data crawler was also developed to collect data that could be analyzed using the Elsevier developer portal (EDP) openAPI. Here, EDP provides the

journal article information, such as the title, abstract, authors, name of the journal, and partial full-text. Our prototype system was designed to run on an Intel Core i7 processor with 32 GB RAM, 256 GB SSD, and Windows 10. A total of 31,748,498 sentences were detected in the 4,200,667 abstracts that were analyzed. When extracting sentences, the SD created a sentence detection model based on a total of 6,282 sentences that were collected manually during the various tests. The number of iterations was set to 20,000 and the cutoff was set to 0. We extract tag patterns of 1 to 5-grams. The clustering data that contain numeric values or invalid characters for the analysis were excluded from the analysis. **Table 5** shows the results of the tag patterns.

Table 5. Result of n-gram tag patterns

n-gram	# of patterns	#of sentences
1	305,674	77,669,865
2	3,370,278	167,368,215
3	10,724,379	132,812,162
4	18,138,429	100,808,987
5	21,850,340	73,234,537

As a result of various analyses, patterns with n>=4 can be composed of a combination of patterns from 1 to 3. Therefore, it was excluded from the analysis of this study. We performed an analysis of n <= 3 and created basic models. **Table 6** shows the results of the analysis by using our model with tagged-entity that can be training data for NE.

Table 6. Sample of NE training data

Tagged-sentence for NE training
<START:entity>Trabzon<END> that is a <START:property>coastal city<END>
<START:entity>Raspberry Pi<END> that is a <START:property>low-cost embedded computer<END>
<START:entity>fine-textured waste material<END> that is a <START:property>mining by-product<END>
<START:entity>weight management<END> was associated with <START:property>better diet quality<END>
<START:method>magnetic separation experiment<END> were conducted with <START:value|property>polyvinyl alcohol<END>

We developed a prototype system to test our model and recruited 5 volunteers for the evaluation. Their characteristics are summarized below:

One Ph.D. with an average of 2 years of experience in english education
One Ph.D. with an average of 4 years of experience in programming
Three graduate students with an average of 1 year of experience in programming

The volunteers were not familiar with this system. However, the system was not inconvenient for users, because it had an interface similar to those of other search systems. The purpose of this study is to analyze a scientific document. Therefore, we limited the keyword sentences for the evaluation to only the abstracts of academic papers. The experiment was repeated 10 times to evaluate the satisfaction. We collected answers to the following research question on this system:

- RQ: Is the relationship between the resulting objects appropriate?

Table 7 describes the validity of search result based on the 4-point scale with the criteria.

Table 7. Validation Criteria of the search result

Score	Designation	Description
4	very satisfied	The search result clearly shows the relations, entities and the properties.
3	Satisfied	The search result only shows one of the relations, entities or the properties.
2	Slightly satisfied	The search result shows the relations, entities or the properties, but it is inappropriate.
1	not satisfied	The search result shows nothing, but the sentence has some of the relations, entities or the properties.

This research question is intended to check if the entity extraction models are appropriate. All subjects answered the appropriateness of the search result based on the evaluation criteria. The mean of the scale point was 2.4. It was lower than we expected. There are several possible explanations for this result. For the example sentence “methods for treating mammalian subjects such as human subjects, having hyperproliferative disorders”, the analysis model extracts “methods for treating mammalian subject” as “method (algorithm)” and “such as human subjects, having hyperproliferative disorders” as “property” respectively. However, for the sentence “particular has emerged as a key strategy for a successful design.”, the model detected the words “particular” and “key strategy for a successful design” have meaningful properties because “particular” exists before the tag “VBG VBN”. We think that is inevitable exceptions because the system handles a natural language. We are currently collecting more data and updating our analytical models to improve quality and performance.

The purpose of this study is to find the entity search, the relationship between the entities, and the FK pattern that can aid extract. The analysis was performed using only the attributes of basic POS. However, to further refine the characteristics of entities existing between FK, it is necessary to define the POS of the word in detail. For instance, the word “after” can be used as a conjunction, a preposition, and an adverb. If it is followed by a noun, such as “I went for a swim after lunch,” it can be a preposition. Additionally, if it connects two clauses, such as “After you’d finished, we completed more tasks,” it can be a subordinating conjunction. However, most POS tagging tools currently process "after" as

"IN." We believe and expect that the analytical performance will be improved if we separate it into “PRPS,” “ADVB,” and “SCNJ” in detail.

5. Conclusions

NER is widely used in the field of natural language processing and is a key element in data mining applications. In this study, a functional-keyword oriented method was proposed that combines stop-words for NE extraction. An analysis method was also proposed based on FK, which consists of identifying stop-word combinations instead of existing noun-based feature analysis methods. While stop-words are generally removed from the preprocessing phase of natural language processing, they are an important factor in this study. The text analyzed in this study was separated into single sentences for POS tagging. Based on the previous results, statistical analyses, such as linguistic and grammatical feature analysis, word frequency, attribute, and classification, were performed. FKs were generated using the analyzed results. Finally, we generated an entity analysis model suitable for extracting entities, such as techniques, methods, and properties, or entity candidates in sentences. Furthermore, the relationships between entities were determined based on the FKs that had collocated words among the non-entity type tags. In future work, an information extraction method will be proposed based on the developed model, along with a complete QA system.

References

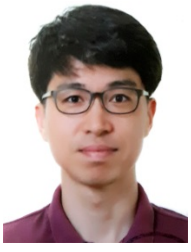
[1] B. Mohit, “Named Entity Recognition,” *Natural Language Processing of Semitic Languages*, pp. 221-245, 2014. Article (CrossRef Link)
[2] J. Cowie, W. Lehnert, “Information extraction,” *Communications of the ACM*, Vol 39, issue. 1, pp. 80-91, 1996. Article (CrossRef Link)
[3] H. LeHong, J. Fenn, *Hype Cycle for Emerging Technologies*, Gartner, 2013.
[4] D. Nadeau, S. Sekine. “A survey of named entity recognition and classification,” *Linguisticae Investigationes*, vol. 30, issue. 1, pp. 3-26, 2007. Article (CrossRef Link)
[5] A. Ritter, S. Clark and O. Etzioni, "Named entity recognition in tweets: an experimental study," in *Proc. of the conference on empirical methods in natural language processing*, pp. 1524-1534, 2011. Article (CrossRef Link)
[6] Y. Choi and J. Cha, "Korean Named Entity Recognition and Classification using Word Embedding Features," *Journal of KIISE*, vol. 43, issue. 6, pp. 678-685, 2016. https://dx.doi.org/doi:10.5626/JOK.2016.43.6.678
[7] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification*, Wiley Interscience, pp. 526-528, USA, 2001. Article (CrossRef Link)
[8] Y. Lu, D. Ji, X. Yao, X. Wei and X. Liang, "CHEMDNER system with mixed conditional random fields and multi-scale word clustering," *Journal of cheminformatics*, vol. 7, issue. 1, 2016. Article (CrossRef Link)
[9] X. Wang, C. Yang and R. Guan, "A comparative study for biomedical named entity recognition," *International Journal of Machine Learning and Cybernetics*, vol. 9, issue. 3, pp. 373-382, 2018. Article (CrossRef Link)

[10] O. Bender, F. J. Och and H. Ney," Maximum entropy models for named entity recognition," in *Proc. of the seventh conference on Natural language learning at HLT-NAACL*, vol. 4, pp. 148-151, 2003. Article (CrossRef Link)
[11] G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis and C. D. Spyropoulos, "Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods," in *Proc. of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 128-135, 2000. Article (CrossRef Link)
[12] N. Chinchor and P. Robinson, "MUC-7 named entity task definition," in *Proc. of the 7th Conference on Message Understanding*, 1997. Article (CrossRef Link)
[13] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato and J. M. Gómez-Berbís, "Named entity recognition: fallacies, challenges and opportunities," *Computer Standards and Interfaces*, vol. 35, issue. 5, pp. 482-489, 2013. Article (CrossRef Link)
[14] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta and F. Ciravegna, "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," *Web Semantics: science, services and agents on the World Wide Web*, vol. 4, issue. 1, pp. 14-28, 2006. Article (CrossRef Link)
[15] L. Reeve and H. Han, "Survey of semantic annotation platforms," in *Proc. of the 2005 ACM symposium on Applied computing*, pp. 1634-1638, 2005. Article (CrossRef Link)
[16] R. Srihari, W. Li, "Information extraction supported question answering," in *Proc. of the 8th Text REtrieval Conference (TREC-8)*, no. 500, pp. 185–196, 2000. Article (CrossRef Link)
[17] A. Maedche and S. Staab, "Ontology learning for the semantic web," *IEEE Intelligent systems*, vol. 16, issue. 2, pp. 72-79, 2001. Article (CrossRef Link)
[18] O. Etzioni, M. Cafarella, D. Downey, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld and A. Yates, "Unsupervised named-entity extraction from the web: An experimental study," *Artificial intelligence*, vol. 165, issue. 1, pp. 91-134, 2005. Article (CrossRef Link)
[19] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, issue. 1-2, pp. 1-135, 2008. Article (CrossRef Link)
[20] A. M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, vol. 5, pp. 339-346, 2005. Article (CrossRef Link)
[21] B. Babych and A. Hartley, "Improving machine translation quality with automatic named entity recognition," in *Proc. of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pp. 1-8, 2003. Article (CrossRef Link)
[22] Y. Chen, C. Zong and K. Y. Su, "A joint model to identify and align bilingual named entities," *Computational Linguistics*, vol. 39, issue. 2, pp. 229-266, 2013. Article (CrossRef Link)
[23] C. Nobata, S. Sekine, H. Isahara and R. Grishman, "Summarization System Integrated with Named Entity Tagging and IE pattern Discovery," in *Proc. of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pp. 1742-1745, 2002. Article (CrossRef Link)
[24] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori and S. Shah, "Multi-document summarization based on the Yago ontology," *Expert Systems with Applications*, vol. 40, issue. 17, pp. 6976-6984, 2013. Article (CrossRef Link)
[25] M. Hassel, "Exploitation of named entities in automatic text summarization for swedish," *NODALIDA'03–14th Nordic Conference on Computational Linguistics*, pp. 9, 2003. Article (CrossRef Link)

[26] T. H. Cao, T. M. Tang and C. K. Chau, "Text clustering with named entities: a model, experimentation and realization," *Data mining: Foundations and intelligent paradigms*, Springer, pp. 267-287, 2012. Article (CrossRef Link)
[27] J. Zhang, Q. Dang, Y. Lu and S. Sun, "Suffix tree clustering with named entity recognition," in *Proc. of the 2013 International Conference on Cloud Computing and Big Data (CloudCom-Asia)*, pp. 549-556, 2013. Article (CrossRef Link)
[28] T. Mandl and C. Womser-Hacker, "The effect of named entities on effectiveness in cross-language information retrieval evaluation," in *Proc. of the 2005 ACM symposium on Applied computing*, pp. 1059-1064, 2005. Article (CrossRef Link)
[29] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez and T. Declerck, "The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions," *Journal of biomedical informatics*, vol. 46, issue. 5, pp. 914-920, 2013. Article (CrossRef Link)
[30] R. Zhang, M. J. Cairelli, M. Fiszman, G. Rosemblat, H. Kilicoglu, T. C. Rindflesch, S. V. Pakhomov and G. B. Melton, "Using semantic predications to uncover drug–drug interactions in clinical data," *Journal of biomedical informatics*, vol. 49, pp. 134-147, 2014. Article (CrossRef Link)
[31] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard and J. H. Holmes, "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation," *Journal of biomedical informatics*, vol. 44, issue. 6, pp. 989-996, 2011. Article (CrossRef Link)
[32] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak and B. Shen, "Biomedical text mining and its applications in cancer research," *Journal of biomedical informatics*, vol. 46, issue. 2, pp. 200-211, 2013. Article (CrossRef Link)
[33] L. De Bruijn, A. Hasman and J. Arends, "Automatic SNOMED classification—a corpus-based method," *Computer methods and programs in biomedicine*, vol. 54, issue. 1-2, pp. 115-122, 1997. Article (CrossRef Link)
[34] U. Leser and J. Hakenberg, "What makes a gene name? Named entity recognition in the biomedical literature," *Briefings in bioinformatics*, vol. 6, issue. 4, pp. 357-369, 2005. Article (CrossRef Link)
[35] K. Shaalan, "Rule-based approach in Arabic natural language processing," *The International Journal on Information and Communication Technologies*, vol. 3, issue. 3, pp. 11-19, 2010. Article (CrossRef Link)
[36] S. Sarawagi, "Information extraction," *Foundations and Trends in Databases*, vol. 1, issue. 3, pp. 261-377, 2008. Article (CrossRef Link)
[37] D. M. Bikel, S. Miller, R. Schwartz and R. Weischedel," Nymble: a high-performance learning name-finder," in *Proc. of the fifth conference on Applied natural language processing*, pp. 194-201, 1997. Article (CrossRef Link)
[38] Y. Wang, Z. Yu, L. Chen, Y. Chen, Y. Liu, X. Hu and Y. Jiang, "Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study," *Journal of biomedical informatics*, vol. 47, pp. 91-104, 2014. Article (CrossRef Link)
[39] S. K. Saha, S. Narayan, S. Sarkar and P. Mitra, "A composite kernel for named entity recognition," *Pattern Recognition Letters*, vol. 31, issue. 12, pp. 1591-1597, 2010. Article (CrossRef Link)
[40] M. Majumder, U. Barman, R. Prasad, K. Saurabh and S. K. Saha, "A novel technique for name identification from homeopathy diagnosis discussion forum," *Procedia Technology*, vol. 6, pp. 379-386, 2012. Article (CrossRef Link)

[41] J. R. Finkel, T. Grenager and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. of the 43rd annual meeting on association for computational linguistics*, pp. 363-370, 2005. Article (CrossRef Link)
[42] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493-2537, 2011. Article (CrossRef Link)
[43] N. Okazaki, "Crfsuite: a fast implementation of conditional random fields (crfs)," http://www.chokkan.org/software/crfsuite/.
[44] D. B. Nguyen, M. Theobald and G. Weikum, "J-NERD: joint named entity recognition and disambiguation with rich linguistic features," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 215-229, 2016. Article (CrossRef Link)
[45] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, "Neural Architectures for Named Entity Recognition," in *Proc. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260-270, 2016. Article (CrossRef Link)
[46] A. E. Borthwick, *A maximum entropy approach to named entity recognition*, New York University, 1999. Article (CrossRef Link)
[47] S. K. Saha, S. Sarkar and P. Mitra, "Feature selection techniques for maximum entropy based biomedical named entity recognition," *Journal of biomedical informatics*, vol. 42, issue. 5, pp. 905-911, 2009. Article (CrossRef Link)
[48] T. Ek, C. Kirkegaard, H. Jonsson and P. Nugues, "Named entity recognition for short text messages," *Procedia-Social and Behavioral Sciences*, vol. 27, pp. 178-187, 2011. Article (CrossRef Link)
[49] Z. Munkhjargal, G. Bella, A. Chagnaa and F. Giunchiglia, "Named entity recognition for Mongolian language," in *Proc. of the International Conference on Text, Speech, and Dialogue*, pp. 243-251, 2015. Article (CrossRef Link)
[50] S. K. Sienčnik,"Adapting word2vec to named entity recognition," in *Proc. of the 20th nordic conference of computational linguistics*, pp. 239-243, 2015. Article (CrossRef Link)
[51] M. Konkol, T. Brychcín and M. Konopík, "Latent semantics in named entity recognition," *Expert Systems with Applications*, vol. 42, issue. 7, pp. 3470-3479, 2015. Article (CrossRef Link)
[52] L. Li, R. Zhou and D. Huang, "Two-phase biomedical named entity recognition using CRFs," *Computational biology and chemistry*, vol. 33, issue. 4, pp. 334-338, 2009. Article (CrossRef Link)
[53] D. Küçük and A. Yazıcı, "A hybrid named entity recognizer for Turkish," *Expert Systems with Applications*, vol. 39, issue. 3, pp. 2733-2742, 2012. Article (CrossRef Link)
[54] J. P. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357-370, 2016. Article (CrossRef Link)
[55] E. J. Huth, "Structured abstracts for papers reporting clinical trials," *Annals of Internal Medicine*, vol. 106, issue. 4, pp. 626-627, 1987. Article (CrossRef Link)
[56] L. B. Sollaci and M. G. Pereira, "The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey," *Journal of the medical library association*, vol. 92, issue. 3, pp. 364, 2004. Article (CrossRef Link)
[57] A. Ratnaparkhi, "A simple introduction to maximum entropy models for natural language processing," *IRCS Technical Reports Series*, pp. 81, 1997. Article (CrossRef Link)
[58] T. M. Mitchell, *Machine learning. WCB*, McGraw-Hill Boston, 1997.

[59] C. D. Manning, C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*, MIT press, 1999. Article (CrossRef Link)
[60] A. M. Mood, F. A. Graybill and D. C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill Kogakusha, 1974.
[61] S. Nam, S. K. Kim, H. G. Kim, V. Ngo and N. Zong, "Structuralizing biomedical abstracts with discriminative linguistic features," *Computers in biology and medicine*, vol. 79, pp. 276-285, 2016. Article (CrossRef Link)
[62] D. Arthur and V. Sergei, "k-means++: The advantages of careful seeding," *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027-1035, 2007. Article (CrossRef Link)



Sangwon Hwang is a research professor of Artificial Intelligence and BigData Medical Center at the Yonsei University Wonju College of Medicine, Korea. He received his Ph.D in computer science from Yonsei University, Korea, in 2014. His research interests include software engineering, data mining, ontology, code reuse, information extraction, machine learning, and artificial intelligence.

Jang-Eui Hong is a professor of Computer Science Department at the school of Electrical and Computer Engineering, Chungbuk National University, Cheongju, Korea. He received his Ph.D in computer science from KAIST, Korea, in 2001. He served as a research member at ADD (Agency for Defense Development) from 2000 to 2002, and also served as a principal consultant at SolutionLink, Co., Ltd. His research interests include software quality, embedded software architecture, low-energy software model, and software process improvement.

Young-Kwang Nam is a professor with the department of Computer and Telecommunications from Yonsei University. He received his B.S. degree in mathematics from Yonsei University in 1982, and the M.S. and Ph.D. degrees in computer science from the KAIST and Northwestern University, in 1985, 1991, respectively. He was a Senior Researcher at SERI (System Engineering Research Institute). His research areas are Programming Language, Software Engineering, Information Retrieval, Database, and XML