ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

An Efficient Damage Information Extraction from Government Disaster Reports☆

Sungho Shin[1] Seungkyun Hong[2] Sa-Kwang Song[1,2*]

ABSTRACT

One of the purposes of Information Technology (IT) is to support human response to natural and social problems such as natural disasters and spread of disease, and to improve the quality of human life. Recent climate change has happened worldwide, natural disasters threaten the quality of life, and human safety is no longer guaranteed. IT must be able to support tasks related to disaster response, and more importantly, it should be used to predict and minimize future damage. In South Korea, the data related to the damage is checked out by each local government and then federal government aggregates it. This data is included in disaster reports that the federal government discloses by disaster case, but it is difficult to obtain raw data of the damage even for research purposes. In order to obtain data, information extraction may be applied to disaster reports. In the field of information extraction, most of the extraction targets are web documents, commercial reports, SNS text, and so on. There is little research on information extraction for government disaster reports. They are mostly text, but the structure of each sentence is very different from that of news articles and commercial reports. The features of the government disaster report should be carefully considered. In this paper, information extraction method for South Korea government reports in the word format is presented. This method is based on patterns and dictionaries and provides some additional ideas for tokenizing the damage representation of the text. The experiment result is F1 score of 80.2 on the test set. This is close to cutting-edge information extraction performance before applying the recent deep learning algorithms.

☞ keyword : Damage Information, Information Extraction, Government Disaster Report, Damage Property, User-generated Text

1. Introduction

Research on social issues that threaten human safety, such as the spread of natural disasters and diseases, continues to attract interest and we can expect steady progress in this area. Problems such as natural disasters do not occur frequently, but they can lead to death and serious property damage when they happen. Therefore, it is necessary to predict the damage caused by natural disasters such as typhoons and floods, and to develop a response strategy for these disasters. Factors that are needed for prediction include data, models, and systems, and the importance of data among them is now increasing. Weather data needed to predict natural disasters can be automatically collected through observation systems such as Automatic Weather Station (AWS) or weather sensors, but the damage data after a disaster is not systematically collected or appropriately shared.

In Korea, data collected during each disaster is recorded and managed by the local government and the central government, and is shared with the public in a report format that describes the damage, along with plain text, rather than raw data. Therefore, information extraction needs to be used for obtaining damage data from government reports. Information extraction researches have mostly focused on unstructured documents such as the web, reports, and social networks. A typical documents created by the government have their own features and the information extraction method considering these features must be applied.

This paper extends upon a previous conference publication [1]. In the previous, we presented an information extraction method based on patterns and dictionaries to extract the damage data from government disaster reports. The extensions include model specification and extension, and enhanced discussions and explanations throughout the paper.

1 Decision Support Technology Lab., Korea Institute of Science and Technology Information, Daejeon, Korea
2 Department of Big Data Analysis, Korea University of Science and Technology, Daejeon, Korea
* Corresponding author (imsmallj@kisti.re.kr)
[Received 23 June 2017, Reviewed 10 July 2017(R2 29 August 2017), Accepted 17 October 2017]
☆ This research was supported by Korea Institute of Science and Technology Information (KISTI).
☆ A preliminary version of this paper was presented at ICONI 2016 and was selected as an outstanding paper.

2. Related Work

2.1 Information Extraction

Information extraction is used for gaining structured information from un/semi-structured machine-readable and user-generated documents [2]. There are many sub-tasks such as Named Entity Recognition (NER) and Relation Extraction (RE). NER is a sub-task of extracting information to find and classify named objects of text into pre-defined categories such as person, organization, location, time representation, quantity, monetary value, percentage, and so on. Some researches on NER focus on extracting temporal expressions such as July 16, 2016 and 2016-07-16 from web articles [3, 4]. NER has been implemented using Conditional Random Field (CRF) [5] and Averaged Perceptron (AP) [6]. Many NER studies have recently focused on adding global capabilities. RE task is generally a search or classification of semantic relations references within a set of text or XML documents. Much research on RE has focused on using the Maximum Entropy (ME) and Support Vector Machine (SVM). There are also many rule-based information extraction systems that can be used to build knowledge bases [7, 8]. They were used in the extraction process from the start of information extraction studies, but some systems recently used merge rules and machine learning methods.

Recently, the emergence of a new paradigm of end-to-end reinforcement learning of deep learning has radically changed the traditional methods of information extraction [9]. The conventional information extraction has been used to be only partially optimized by applying appropriate methodology to each extraction step. End-to-end reinforcement learning is a method of optimizing the whole step of information extraction. Once the input data and the final output data are provided as training data, the algorithm optimizes parameters to produce the best results. Although the emergence of these new technologies seems to make all tasks possible, there is a drawback that it is difficult to expect high accuracy if sufficient training data can not be obtained. End-to-end reinforcement learning can be applied for this study, but it was difficult to apply the latest technology due to insufficient disaster history data.

2.2 Various Types of User-generated Text

There are variety of user-generated texts such as web documents, market analysis reports, blogs, tweets, posts, and CFP (in academic conferences)[10]. The structure of user-generated text affects highly on the complexity of the extraction [11, 12]. Generally, if the text is written according to a basic sentence structure, the result of syntactic analysis is highly accurate, so information extraction is relatively easy.

However, if the components of a sentence such as subject, object, verb, etc. are omitted or there are incomplete sentences, extraction via traditional information extraction methods results in low accuracy. In the field of information extraction, these text data are defined as noisy user-generated text, and the extraction is being studied as a separate field of study. Table 1 shows categories of text and characteristics of each.

(Table 1) Categories of Text and Characteristics

Categories of Text	Characteristics
Web document, news article, academic paper, market analysis report	Well-written according to grammar
Tweet, post	Special characters and icons (!!!, ???, ♥♥♥, Oooooops etc.)
Poet, lyric	Frequent irony, metaphor and paradox
Government report (in South Korea)	Highly summarized and POS omitted

3. Background

3.1 Disaster Damage Information

When natural disasters such as typhoons and floods occur, they are usually accompanied by damage. Damage refers to deaths and injuries to human bodies, loss of property, and reputationdecrease. There are variety of properties for damage such as the source, target, type, volume, unit, location, and time.

(Table 2) Damage Properties

Categories of Text	Damage Target	Damage Unit	Damage Type
Typhoon Rainfall Earthquake Landslide Forest Fire Drought	Person	명 (inKorean)	Death
			Injury
			Missing
	Building	동 (in Korean)	Collapse
			Flooding
			Washed Away
	Vehicle	대 (in Korean)	Flooding
			Washed Away
	Bridge	개 (in Korean)	Collapse Flooding

The cause of damage is disaster such as rainfall, typhoon, and earthquake, and the target includes objects that are affected by the disaster. The type of damage varies depending on the damage target. For example, if the target is a person, the type must be death, injury, or missing. If the target is a building, the type has the types of collapse, flooding, or washed away. The damage unit depends on the target, and can be '명', '동', '대', '개'(in Korean), and so on. Different with English, Korean has designated expressions for the units themselves. They are used in mentioning the number of objects, coming after the number. For example, 33 people in English is correspond to 33명 in Korean. Even though people is same as 사람 in Korean, Korean use 33명 in counting people instead of 33사람. That is why we define the unit as one of properties. Table 2 shows examples of some damage properties.

3.2 Disaster Damage Report

Damage caused by a disaster is first checked out by the local government after the disaster has occurred. The results of the investigation are aggregated and managed by the central government [13]. The government only publicly releases the report pertainingto province/city damage aggregated through this process for each disaster. To predict disaster damage, the data for each hour or day and (at least) each region are required. The reason for this is that incremental predictions according to the passage of time are needed and regional (or district or neighborhood) predictions must be made so that residents of the regions escape before having damages.

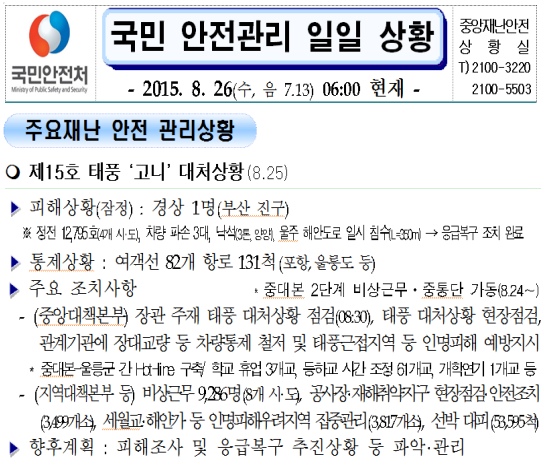
The damage data for a disaster is created by the government in report format, which includes text, as shown in the image. In order to obtain the raw data, information extraction task is needed to extract the damage information from the text. Figure 1 shows an example of the government damage report. The characteristics of the document are as follows.

First, the sentences are shorter than normal sentences in other documents because subjects and predicates areomit ted.Forexample,thetextinredrectangleinthefigureistranslate dintoEnglishasfollows.Belowisnotatypicalsentence.

The state of damage (tentatively): 1 slightly wounded person (Jin districtinPusan).

Second, each piece of damage is separated by a comma. Comma means the end of a damage. After a space, another damage starts.

Third, parenthesesare used for describing sub-damages or properties of the damage. There are several parentheses in the above document. Some contain the properties in-between and the others are selected for sub damage. For example, if the number of slight wounded people is 3, the sentence can be 3 slightly wounded people (1 Seoul, 2 Pusan).


국민안전처 Ministry of Public Safety and Security

국민 안전관리 일일 상황

- 2015. 8. 26(수, 음 7.13) 06:00 현재 -

중앙재난안전상황실 T) 2100-3220 2100-5503

주요재난 안전 관리상황

❍ 제15호 태풍 '고니' 대처상황(8.25)

▸ 피해상황(잠정) : 경상 1명(부산 진구)

※ 정전 12,705호(4개 시·도), 차량 파손 3대, 낙석(3톤, 양맥, 울주 해안도로 일시 침수(L=30m) → 응급복구 조치 완료

▸ 통제상황 : 여객선 82개 항로 131척(포항, 울릉도 등)

▸ 주요 조치사항 ◦ 중대본 2단계 비상근무 · 중통단 가동(8.24~)

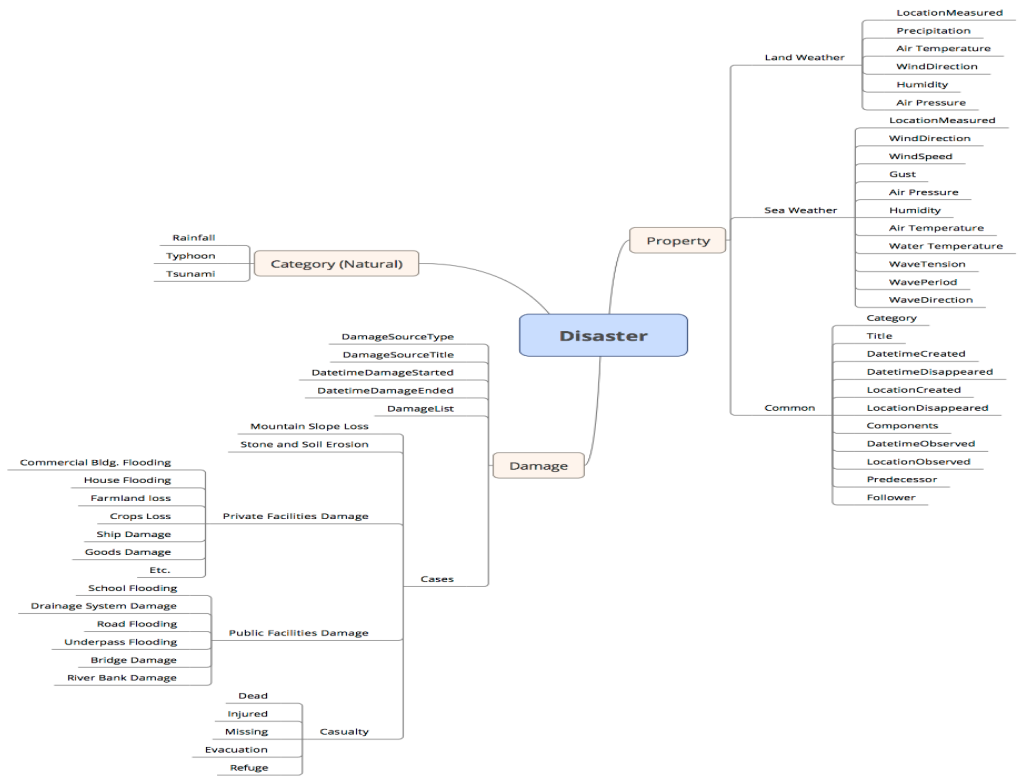
- (중앙대책본부) 장관 주재 태풍 대처상황 점검(08:30), 태풍 대처상황 현장점검, 관계기관에 장대교량 등 차량통제 철저 및 태풍근접지역 등 인명피해 예방지시

◦ 중대본-울릉군 간 Hot-line 구축(학교 휴업 3개교, 등하교 시간 조정 6개교, 개학연기 1개교 등

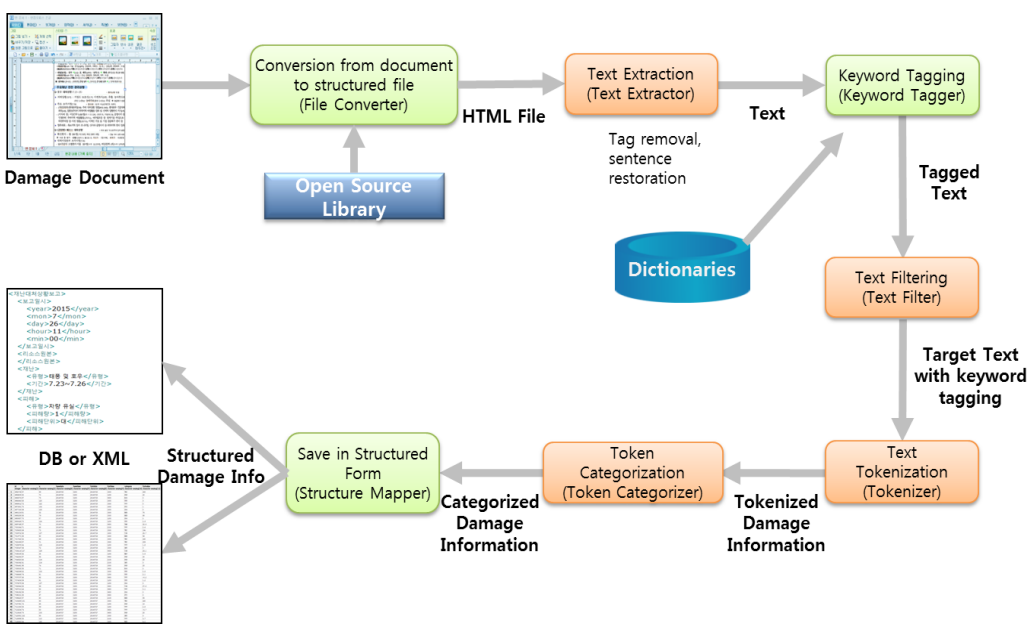
- (지역대책본부 등) 비상근무 9,286명(8개 시·도), 공사장·재해취약지구 현장점검·안전조치(3,499개소), 세월교·해안가 등 인명피해우려지역 집중관리(3,817개소), 선박 대피(53,595척)

▸ 향후계획 : 피해조사 및 응급복구 추진상황 등 파악·관리

(Figure 1) Example of government report



(Figure 2) Data modeling



(Figure 3) Process of information extraction from government report

4. Research Design

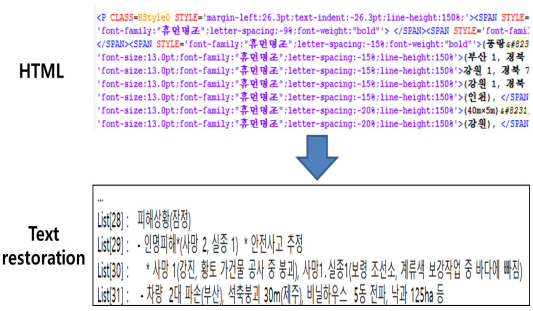
4.1 Data Modeling

We first perform data modeling to process texts that appear in government disaster reports. Data modeling is to arrange the properties of the disaster and damage data and find relationships between these properties. For example, a disaster can be described in terms of its category and properties and the accompanying damage, as well as a list of documents that include disaster information. Furthermore, detailed properties for damage can be defined such as the damage category, history, and type [14]. Figure 2 shows a diagram of the relationships between this kind of information.

4.2 Information Extraction Process

Figure 3 shows a diagram of the information extraction process. First, because the government damage report is stored in a WORD file format rather than a TXT or XML file that can be easily text-processed, the task that extracts texts from the WORD file is needed. However, direct extraction from the WORD file is not possible; rather, it can only be extracted after the WORD file has been converted into an HTML or XML file. In this case, the texts are separated into word units such that the texts must be restored after the structure is converted. After text restoration, dictionaries are used to tag key words that we focus on. The keyword-tagged information is used to remove portions of the texts that are not related to damages. After that, it is tokenized according to a piece of damage, and each damage token is placed in to one of three categories: source disaster, damage term, and damage information. The categorized damage tokens are saved as is appropriate for the output data structure based on the category information.

The most important part of this process is the step in which the texts are tokenized according to the piece of damage. Figure 4 shows the result after text restoration in the process. In the html format, each piece of damage information starts with <P> tag. Texts in a same <P> tag is restored into a same list number such as List[28] and List[29]. Each list represents title or damage information or detail explanation.



(Figure 4) Text Restoration

Figure 5 shows the text after keyword tagging. The texts within the angle brackets(<>) are damage information in a same damage group. Damage grouping is done in the part of text extraction. Damage information in a group is needed to be separated in to each damage token. For the task, commas are used to separate each damage token.

```
< 제12호 태풍 '할롤라' 및 호우 대처상황(7.23~25)>
12/number, 호/unit, 태풍/damage_source, 호우/damage_source

< 피해상황(잠정) : 지방도 56호선(인제) 사면 토사(20톤) 유출>
지방도/target, 56/number, 호/unit, 인제/county, 사면/target, 토사/target, 20/number, 톤/unit, 유출/damage

< 농배수로 유실(철원, 10m),>
농배수로/target, 유실/damage, 철원/county, 10/number, m/unit)

< 차량피해 5대(유실1, 침수1, 파손3),>
차량/target, 5/number, 대/unit
                < 유실1 >
                유실/damage, 1/number
                < 침수1 >
                침수/damage, 1/number
                < 파손3 >
                 파손/damage, 3/number

< 주택 19세대(수원7, 용인1, 화성8, 오산2, 김포1) 침수(배수 완료),>
주택/target, 19/number, 세대/unit, 침수/damage
                < 수원7 >
                수원/city, 7/number
                < 용인1 >
                용인/city, 1/number
                …
```

(Figure 5) Result after keyword tagging

Commas are often used for other purposes, so commas are used for tokenizing only if they are followed by another damage token which starts with damage target, damage amount, damage unit, damage type, and so on. For example, the text in figure 6, < *농배수로 유실(철원,10m)* > *and* < *차량피해 5대(유실1, 침수1, 파손3)* > were on the same text line before text tokenization, but they contain commas in the middle so they are separated into individual tokens. The commas can exist within parentheses, but here the commas refer to the properties of damage tokens before the

parentheses or are used for subdividing the damage. To summarize, in order to extract damage information, it is important to separate the damage texts into damage tokens, and to do this, the presence of commas and damage properties is used.

4.3 Input and Output Data Format

As previously mentioned, the input data include government reports consisting of semi-structured sentences. Ultimately, the damage data extracted from text is stored in XML format. Figure 6 shows the format of output data.

```
<source_disaster>  </source_disaster>
<term>  </term>
<damages>
    ...
    <damage>
        <target>  </target>
        <type>    </type>
        <volume>  </volume>
        <unit>    </unit>
        <place>
            <province_city>   </province_city>
            <district_county> </district_county>
            <dong>            </dong>
        </place>
        <time>
            <year>   </year>
            <month>  </month>
            <day>    </day>
        <time>
    </damage>
</damages>
```

(Figure 6) Output data format

4.4 Patterns for Information Extraction

Considering and analyzing the characteristics of texts in the government reports, we built some patterns and applied them for information extraction. Because of the difference of using commas in the text, patterns are respectively built for main damage information and sub damage information. For example, the expression of '차량5 대' in Korean means that five vehicles were damaged. '차량' is a kind of target, '5' points number, and '대' responses to unit. Therefore, the pattern can be *Target Number Unit*. We built dozens of patterns for main damage information and sub damage information respectively.

(Table 3) Patterns for Information Extraction

Patterns for Main Damage Info.
Target (Number Unit) Type Target Type Number Unit Target Type Number Unit (Region) Target Type (Region Number Unit) Target Type (Region, Number Unit) Target Type (Number Unit) Target Type (Number) Target Number Unit Target Number Unit Type …
Patterns for Sub Damage Info.
Region Number Region Number Unit Target Number Target Number Unit Type Number Type Number Unit Target Type Number Target Type Number Unit …

5. Experiment and Result

5.1 Experiment

A damage information extraction system was implemented according to the method described in section 3. The damage reports are managed and released by the Ministry of Public Safety and Security. The reports have been released since 2002, roughly 300 to 400 documents each year. To evaluate the performance of our system, we selected documents that describe the damage caused by top 10 typhoons that have struck South Korea since 2002. The 10 typhoons are RUSA(in 2002), MAEMI (in 2003), NABI (in 2005), NARI (in 2007), KOMPASU (in 2010), MUIFA (in 2011), TEMBIN (in 2012), BOLAVEN (in 2012), SANBA (in 2012), and GONI (in 2015). For each typhoon, we used one final comprehensive document describing the damage status as our extraction target.

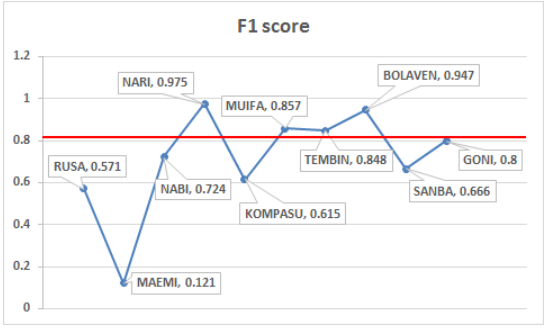
We evaluated the system performance by using the balanced F score (F1 score), which is commonly used in information retrieval/extraction. To calculate the F1 score, the precision and recall values must be respectively calculated. True Positive, False Positive, and False Negative values of each document are counted.

4.2 Result and Implication

As shown in the table 2, the performance of the system resulted in an F1 score of approximately 80.2%(precision 95%, recall 68%), which is close to state-of-the-art information extraction before the application of recent deep learning algorithms.

(Table 6) Experiment Result

Typoon	TP	FP	FN	F1
RUSA	6	1	8	0.571
MAEMI	2	0	29	0.121
NABI	25	2	17	0.724
NARI	99	3	2	0.975
KOMPASU	8	0	10	0.615
MUIFA	9	1	2	0.857
TEMBIN	14	0	5	0.848
BOLAVEN	9	0	1	0.947
SANBA	7	0	7	0.666
GONI	10	1	4	0.800
Total	189	8	85	-
avg. Precision	0.959			
avg. Recall	0.689			
avg. F1 score	0.802			



(Figure7) Comparison of F1 Score by Each Typoon

In Figure 7, the result of typhoon NARI is best and that of MAEMI is worst. The difference is from the sentence structure. The present patterns used in our system can not cover the sentence structure of MAEMI document. We need to add more patterns for the structure of the document. On the other hand, our system is best suitable for the sentence structure of NARI document. Overall, our system is expected to extract disaster damage information from government reports enough to use them for the damage prediction.

In recent years, most of the world's best methods in information extraction are deep running algorithms. In the case of relationship extraction, the world's best accuracy is F1 score of around 86% (year of 2014). Prior to the deep run algorithms, a kernel based method showed the best F1 score by about 80%. The method presented in this study is neither kernel based nor deep learning. However, when choosing a method for practical information extraction in business, we are not recommended to choose algorithms just considering high accuracy. Appropriate methods should be chosen according to the nature of documents to be processed or sentence structure. Disaster reports by South Korea government do not follow the grammar seen in full sentences, but list up the objects of the disaster, the cause, the degree of damage, the place, the time, etc. Therefore, it is possible to develop techniques with minimum effort that can be applied to actual job by analyzing the characteristics of such sentences and applying appropriate rules. In the test, the F1 score is 80.2%, which is acceptable.

6. Conclusion

IT is expected to be used to predict in advance disasters and damage. Obtaining the raw data is one of the most important tasks for prediction. In South Korea, the data related to the damage is checked out first by each local government and then is aggregated by the federal government. This data is included in the federal disaster report by disaster cases, and it is difficult to obtain the raw data even for research purposes. We decided to apply the information extraction method to extract the raw data from the text. Our method is based on patterns and dictionaries, with some additional ideas to consider the nature of government reports and to tokenize each piece of damage expression in text. The accuracy is approximately 80.2 F1-score and is close to cutting-edge information extraction research.

Reference

[1] S. Shin, S. Hong, and S. Song, "Disaster Damage Information Extraction from Government Reports," Proceedings of the 8th International Conference on

Internet (ICONI), 2016.

[2] Kim et al., "The Management of Disaster Information based on Big Data and Cloud Computing," Journal of Disaster Prevention, Vol. 17, no. 2, pp. 14-33, 2015.

[3] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky, "Semeval-2010 task 13: Tempeval-2," In Proc. of the 5th International Workshop on Semantic Evaluation (SemEval'10), pp. 57–62, 2010. http://www.aclweb.org/anthology/S10-1010

[4] J. Lee and Y. Kwon, "A Proposal of Methods for Extracting Temporal Information of History-related Web Document based on Historical Objects Using Machine Learning Techniques," Journal of Internet Computing and Services (JICS), Vol. 16, no. 4, pp. 39-50, 2015.

[5] T. Baldwin, M. Catherine, B. Han, Y.B. Kim, A. Ritter, W. Xu, "Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition," Proceedings of Workshop on Noisy User-generated Text (WNUT), 2015. https://doi.org/10.18653/v1/w15-4319

[6] Information extraction, https://en.wikipedia.org/wiki/Information_extraction

[7] S. G. Small and L. Medsker, "Review of information extraction technologies and applications," Neural Computing and Applications, Vol. 25, no. 3, pp. 533-548, 2014. https://doi.org/10.1007/s00521-013-1516-6

[8] S. Shin, H. Jung, and M. Y. Yi, "Building a Business Knowledge Base by a Supervised Learning and Rule-based Method," KSII Transactions on Internet and Information Systems, Vol. 9, no. 1, pp. 407-420, 2015. https://doi.org/10.3837/tiis.2015.01.025

[9] M. Miwa and M. Bansal, "End-to-End Relation Extraction Using LSTMs on Sequences and Tree Structures," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1105–1116, 2016. https://doi.org/10.18653/v1/p16-1105

[10] H. Yoon, J. Kim, J. Park, and T. Chang, "Development of Electronic Documents and Management System for Transfer of Disaster Damage and Recovery Information," Journal of Society for e-Business Studies, Vol. 20, no. 2, pp. 15-26, 2015. https://doi.org/10.7838/jsebs.2015.20.2.015

[11] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, features Induction and Web-Enhanced Lexicons," Proceedings of Conference on Computational Natural Language Learning, pp. 188-191, 2003. https://doi.org/10.3115/1119176.1119206

[12] M. Collins, "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms," Proceedings of Empirical Methods in Natural Language Processing, pp. 1-8, 2002. https://doi.org/10.3115/1118693.1118694

[13] S. Shin, C. H. Jeong, D. Seo, S. P. Choi, and H. Jung, "Improvement of the Performance in Rule-Based Knowledge Extraction by Modifying Rules," Proceedings of the 2nd International Workshop on Semantic Web-based Computer Intelligence with Big-data, 2013. http://inscite.kisti.re.kr/cfp/SWCIB2013/proceeding/swcib2013_submission_3.pdf

[14] C. N. Seon, J. H. Yoo, H. Kim, J. H. Kim, and J. Seo, "Lightweight Named Entity Extraction for Korean Short Message Service Text," KSII Transactions on Internet & Information, Vol. 5, no. 3, pp. 560-574, 2011. https://doi.org/10.3837/tiis.2011.03.006

◐ Authors ◑

Sungho Shin

2000 B.S in Business Administration, Kyungpook National Univ., Daegu, Korea
2002 M.S in Management Information Systems, Kyungpook National Univ., Daegu, Korea
2015 Ph.D candidate in Knowledge Service Engineering, KAIST, Daejeon, Korea
2002~present: Senior Researcher, KISTI, Daejeon, Korea
Research Interests: Information Extraction, Big Data, Artificial Intelligence, Deep Learning
E-mail : maximus74@kisti.re.kr

Seungkyun Hong

2014 B.E in Microsoft IT, Keimyung Adams College, Daegu, Korea
2014~present Ph.D student (Integrative) in Big Data Science, UST, Daejeon, Korea
Research Interests: Deep Learning, Machine Learning, Big Data, etc.
E-mail : xo@kisti.re.kr

Sa-Kwang Song

1997 B.S in Statistics, Chungnam National Univ., Daejeon, Korea
1999 M.S in Computer Science, Chungnam National Univ., Daejeon, Korea
2010 Ph.D in Computer Science, KAIST, Daejeon, Korea
2014~Present: Professor, Dept. of Big Data Science, University of Science and Technology, Daejeon, Korea
1999~2000: Researcher, Dept. of Information Retrieval, ETRI, Daejeon, Korea
2000~2003: Team Leader, Dept. of Information Retrieval, SearchCast Inc., Seoul, Korea
2005~2010: Senior Researcher, Dept. of Bioinformatics, ETRI, Daejeon, Korea
2010～present: Principal Researcher/Dept. Head, Decision Support Technology Research Lab., KISTI, Daejeon, Korea
Research Interests: Big Data, Machine Learning, Text Mining, Natural Language Processing, etc.
E-mail : esmallj@kisti.re.kr