Proceedings of the Korean Nuclear Society Spring Meeting

Kwangju, Korea, May 2002

A Methodology for Improving the SIS-RT in Analyzing the Traceability of the Documents written in Korean Language

Yeong Jae Yoo, Man Cheol Kim, Poong Hyun Seong

Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu
Daejon, Korea 305-701

Abstract

Inspection is widely believed to be an effective software verification and validation (V&V) method. However, software inspection is labor-intensive. This labor-intensive nature is compounded by a view that since software inspection uses little technology, they do not fit in well with a more technology-oriented development environment. Nevertheless, software inspection is gaining in popularity. The researchers of KAIST I&C laboratory developed the software tool managing and supporting inspection tasks, named “SIS-RT.” SIS-RT is designed to partially automate the software inspection processes. SIS-RT supports the analyses of traceability between the spec documents. To make SIS-RT prepared for the spec document written in Korean language, certain techniques in natural language processing have been reviewed. Among those, the case grammar is most suitable for the analyses of Korean language. In this paper, the methodology for analyzing the traceability between spec documents written in Korean language will be proposed based on the case grammar.

1. Introduction

The software used in the NPP protection systems is required to have extremely high reliability. In order to produce software of high reliability, very rigorous V&V activities shall be performed throughout the entire life cycle, and generally, it is impossible to verify that the software has aimed reliability through only a conventional testing method. Software inspection technique is widely used in industrial fields.

The technique is considered to be more powerful than the technique with only testing because it can catch out the errors of software to be developed in the early stage of the life cycle. But it has such shortcomings that it is labor-intensive and the rigor of the process cannot be guaranteed because the human inspector mainly performs it. Therefore, the researchers of KAIST I&C laboratory developed the software tool managing and supporting

inspection tasks, named “SIS-RT.”

SIS-RT has two main functions: One is the analysis and rearrangement of the spec documents based on checklist. The other is the analysis of traceability between two spec documents.

In the early stage, the traceability analysis tool of SIS-RT merely displays the sentences of documents in the matrix form, therefore the inspector must make a personal inspection for the analysis. But there are some demands for the improvements of that point, SIS-RT has been upgraded to calculate and display the similarities between the sentences and now it is under the improvements of user interfaces.

But, it is planned to make the documents in Korean language for the future projects, thus SIS-RT is required to have the ability to analyze the traceability between the documents written in Korean language.

In addition, Korean language has different features compared with English in the viewpoint of linguistics and the sentences in the spec documents for NPP protection systems are limited in their forms compared with the sentences of living language. Having an eye on these points, we started this research to set up the methodology for improving the capability of SIS-RT for the traceability analysis.

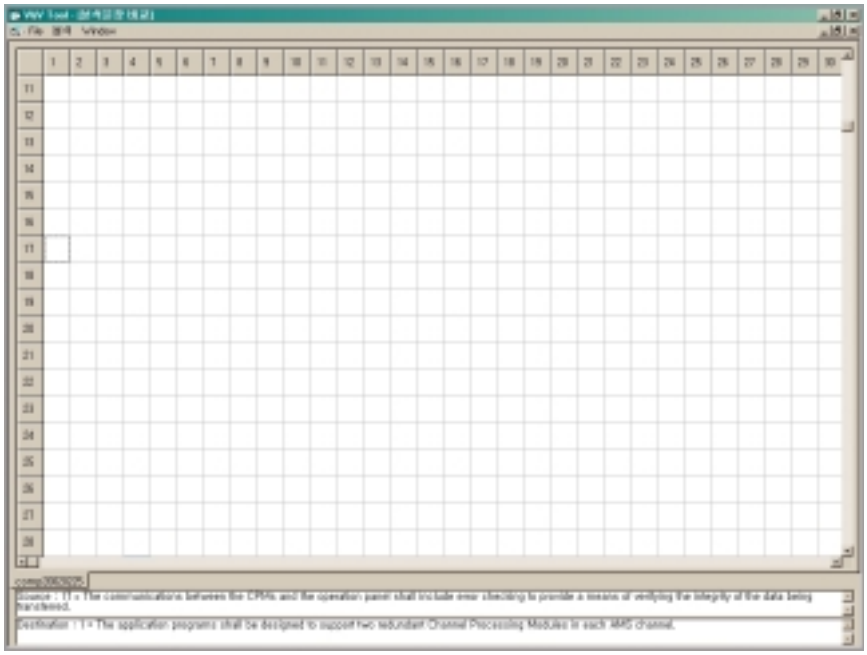

Fig. 1 Traceability Analyses window of SIS-RT (early stage)

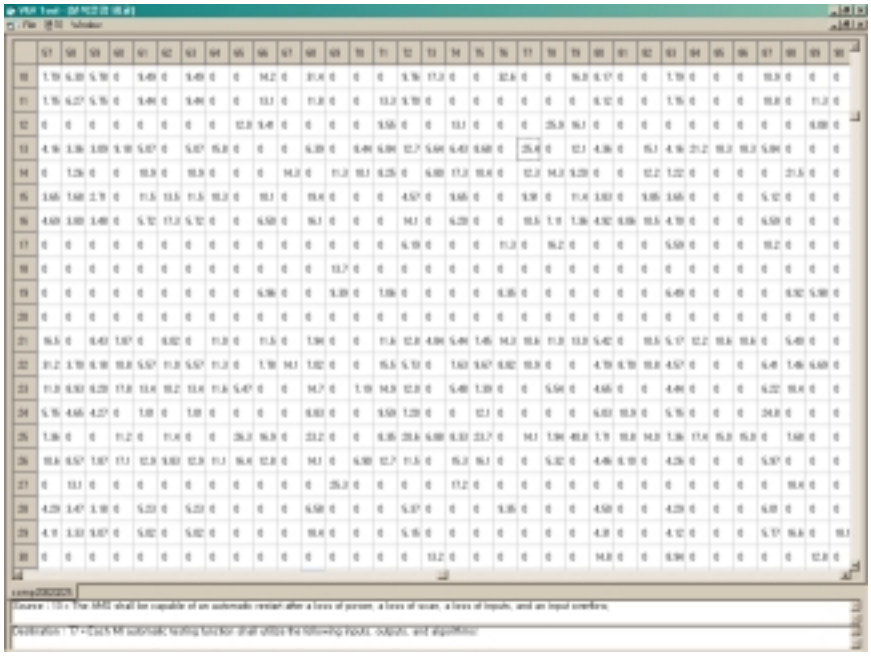

Fig. 2 Traceability Analyses window of SIS-RT (improved)

2. Related Works

Differ from English, Korean language has free word order (scrambling) and inflections. And there are many ellipses of the essential parts. Furthermore, it has agglutinative characteristics that a word is formed with an essential morpheme and a formal morpheme.

As a consequence, linguistic methods are more favored than the statistical methods for the analysis and processing of Korean language. Linguistic methods modify and simplify the grammars of natural languages, then analyze the language. There are typical grammars as follows:

Phase structure grammar (Chomsky)

This grammar was frequently used in the early systems for the simple and efficient parsing method. But it can not represent the general phenomena of Korean language such as omission or free word order and its rules are too complex.

Unification grammar

This describes the grammar by the features of words and analyzes the sentence by the unifications of these features. It describes the omissions and free word orders well. But it requires the dictionary of the features and makes the system slow.

Dependency grammar (Tesniere)

Dependency grammar converts the sentence structure into the dependency relations. It well describes free word orders and omissions because it only computes the relations between words.

Case grammar (Fillmore)

This grammar focuses on the semantics of a language. It does not adhere to the formality of the language, thus it is suitable for the analyses of Korean language

Among these, dependency grammar and case grammar are favored for their suitability to scrambling and omitting. There are many studies about the natural language processing system using these grammars especially in the department of computer science, KAIST.

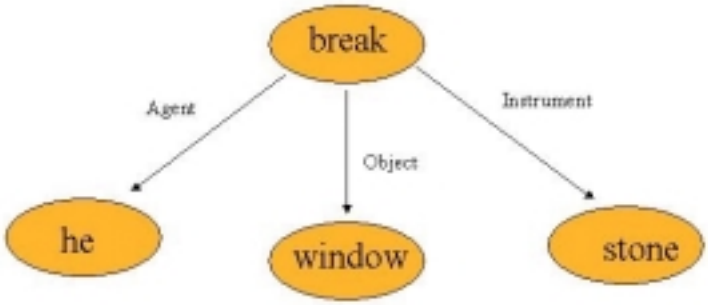

Fig. 3 Case analysis of the sentence “He broke the window with a stone.” [4]

3. Strategy for Analyzing the Traceability of Sentences in the Specification Documents Written in Korean Language

In this research, a methodology for analyzing the traceability will be proposed mainly based on the concepts of case grammar. As SIS-RT traceability analysis tool accepts sentences as inputs, the bases of the approach will be the comparison between tow sentences.

In the upgrading of the tool for English document, the similarity merely computed based on the number of words coexisting in the two sentences. But in Korean language, it is possible to grasp the cases of the substantives with the information about the case frames of verbs and the particles added to the substantives, therefore the substantives of the same cases can be compared to get the similarity of two sentences.

The procedure to get the similarity being embodied in SIS-RT as follows:

Distinction of the important substantives using the corpus
Discernment of surface cases using the information of particles
Separation of the stem of declinable words though the morphological analyses
Discernment of deep cases with the information of the case frames
Calculation of the similarity by comparing the substantives of the same cases

3.1 The Corpus

The proposition core of a simple sentence is composed with one or more substantives

and a predicate. [Nam 1978] Thus, for the purpose of analyzing the case structure, substantives must be extracted from the sentence.

In the sentences of the spec document for a NPP protection system, substantives are usually nouns: bistable processor, coincidence logic, signal, cable, etc. Such nouns are limited in their amounts compared with them in the living language, the corpus of them can be easily established. Once the corpus is made, the essential nouns (substantives) can be extracted from the input sentences.

3.2 Case Particles (Particles that represent the cases of substantives)

In Korean language, a noun can be followed by an auxiliary verb, a suffix or a particle. As the particles are limited in their numbers, it can be easily analyzed and processed. If the sufficient numbers of sentences are given, it is possible to classify the categories of case particles and the cases they represent. After such analyses, the surface cases of the essential nouns can be determined.

Table 1. Requisite cases of Korean language [3]

일련번호	이름	일련번호	이름
1	AGT(agent)	8	MEA(means)
2	OBJ(object)	9	TIM(time)
3	TAR(target)	10	PRT(participant)
4	LOC(location)	11	FCS(focus)
5	INS(instrument)	12	SOR(source)
6	EXP(experience)	13	ELM(element)
7	REA(reason)		

Table 2. Case determining postpositions [3]

일련번호	격판별 대표조사	일련번호	격판별 대표조사
1	이, 가, 은, 는, 께서, 에게서	5	에, 에게
2	을, 를	6	로, 로써, 으로서
3	와, 과	7	로, 으로
4	에, 에서	8	로부터, 한테, 한테서, 에게서

3.3 Morphological Analyses

The surface cases of nouns can be determined from case particles, but the deep cases can be determined from the case frames of verbs. In Korean language, inflections of verbs are quite frequent. Therefore the stem of a verb must be extracted by morphological analyses in order to use the information about case frames of the verb.

For the general morphological analyses, it should be analyzed for all morphemes which morphemes can be added on the right or left side of them. This work is exhaustive, but the subjective documents of SIS-RT are specified for the NPP control and protection systems. Thus it is planned to classify the morphemes attached to the verbs and the patterns of inflections of them through analyzing sample sentences.

3.4 Case Frames

A verb has one or more case frames. The case frame is the frame that represents the deep cases of the substantives in a sentence and the lists of the case particles determining the deep cases. One verb can have different case frame according to the morphemes added to it, and there exist some rules. So the analyses about case frames are easy.

Once the case frame of a verb is determined, the deep cases of substantives can be determined. Then comparing the nouns of same (deep) cases in two sentences, it is expected to achieve a more detailed traceability analysis.

날다	(1	((2	1)))			
날뛰다	(1	((2	1)))			
날리다	(1	((1	1)(2	2)(4	4)))

Fig. 4 Verbal case frame [3]

4. Conclusions & Future Works

In this research, applying the concept of case grammar, the methodology and the procedure for improving the performance of SIS-RT tool were proposed. The essential point is that the analyzed documents are specified in the nuclear domain, thus the rules required for the analyses will be relatively simple and it is expected to produce a system performing satisfactorily in a practical respect. For promoting the efficiency and the accuracy of analyses, the rules of case grammar focused on the sentences in the nuclear field and they are under setting based on the plenty amount of sample sentences.

References

1. Jae-Woo Kang, “A Design and Implementation of Hangul Spelling and Word-spacing Checker using Connectivity Information”, M.S. Thesis, Department of Computer Science, KAIST, 1990

2. Dong-Un, An, “A Corpus-based Modality Generation for Korean Predicates”, Ph.D. Thesis, Department of Computer Science, KAIST, 1995

3. Hyeon-Sung, Han, “A Design and Implementation of Automatic Indexing by using Syntactic Analysis for Korean Text”, M.S. Thesis, Department of Computer Science, KAIST, 1991

4. Young-Rim, Choi, “Implementation of a Korean Case Analyzer using Neural Networks”, M.S. Thesis, Department of Computer Science, KAIST, 1994

5. Chung-Won, Seo, “Dependency Parsing of simple Korean Sentence using Verb Case frame”, M.S. Thesis, Department of Computer Science, KAIST, 2000

6. Dong-Un, An, “Transforming Morphemes into Sentence Constituents in Analyzing Korean Language for Machine Translation”, M.S. Thesis, Department of Computer Science, KAIST, 1987

7. Gil-Bae, Yoon, “A Study on the Case Classification of Natural Language – with Regard to the Sentences of Computer Science Literatures”, M.S. Thesis, Department of Computer Science, KAIST, 1986

8. Jae-Hoon, Kim, “Construction of Korean Case Frames for Generation of Korean Case postposition in the Interlingual Machine Translation”, M.S, Thesis, Department of Computer Science, KAIST, 1988

9. Makoto Nagao, “Natural Language Processing”, Hong Reung Science Press Inc., 1998