Journal of the Korean Nuclear Society
Volume 36, Number 4, pp.295 303, August, 2004

Development of a Traceability Analysis Method Based on Case Grammar for NPP Requirement Documents Written in Korean Language

Yeong Jae Yoo, Poong Hyun Seong, and Man Cheol Kim
Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejeon, Korea 305-701
yjyoo@bnftech.com

(Received September 5, 2003)

Abstract

Software inspection is widely believed to be an effective method for software verification and validation (V&V). However, software inspection is labor-intensive and, since it uses little technology, software inspection is viewed upon as unsuitable for a more technology-oriented development environment. Nevertheless, software inspection is gaining in popularity. KAIST Nuclear I&C and Information Engineering Laboratory (NICIEL) has developed software management and inspection support tools, collectively named "SIS-RT." SIS-RT is designed to partially automate the software inspection processes. SIS-RT supports the analyses of traceability between a given set of specification documents. To make SIS-RT compatible for documents written in Korean, certain techniques in natural language processing have been studied [9]. Among the techniques considered, case grammar is most suitable for analyses of the Korean language [3]. In this paper, we propose a methodology that uses a case grammar approach to analyze the traceability between documents written in Korean. A discussion regarding some examples of such an analysis will follow.

Key Words : traceability, case grammar, linguistic analysis, Korean language, natural language processing

1. Introduction

The software used in NPP protection systems must be extremely reliable. In order to produce highly reliable software, rigorous V&V activities must be performed throughout the entire software life cycle. Generally, it is impossible to verify that a given software program has achieved reliability by conventional testing methods alone. Therefore, software inspection is widely used in industries that require highly reliable software.

The use of software inspection is considered a more powerful means of verification than the use of alone, because software inspection can detect errors that may develop in the early stages of a software program's life cycle. However, software

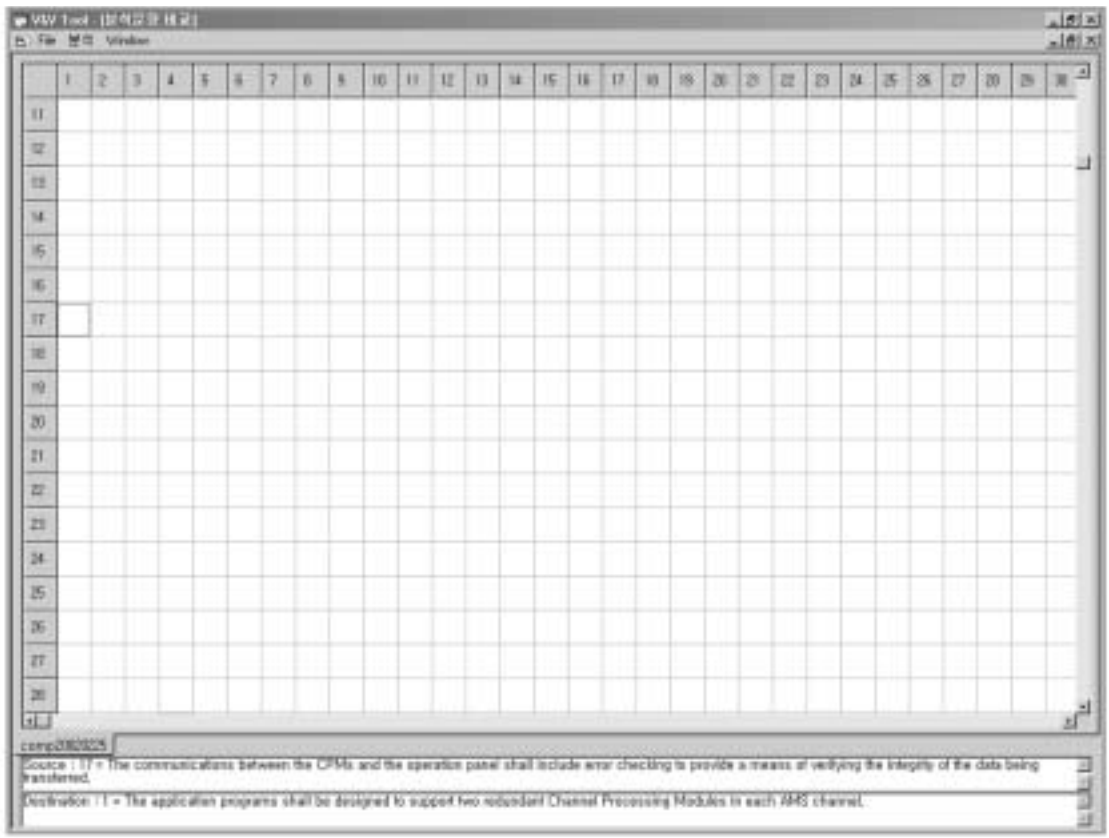

Fig. 1. Traceability Analyses Window of SIS-RT (early stage)

inspection also has shortcomings: it is labor-intensive, the inspections are potentially subject to human error, and the rigor of the process cannot be guaranteed. To address these drawbacks in software inspection, KAIST Nuclear I&C and Information Engineering Laboratory (NICIEL) has developed SIS-RT, a software tool for managing and supporting the software inspection process.

SIS-RT has two main functions: (1) to analyze and rearrange spec documents based on a checklist, and (2) to analyze the traceability between two spec documents.

In the early stages of SIS-RT, the traceability analysis tool of SIS-RT merely displays the sentences of documents in a matrix form; therefore, the inspector must conduct a personal inspection for the analysis (Fig. 1-1). In a recent improvement, SIS-RT has been upgraded to calculate and display similarities between sentences and now (Fig. 1-2).

Thus, SIS-RT must have the ability to analyze the traceability between documents written in Korean. In addition, the Korean language has different linguistic features from English, and the sentences in spec documents for NPP protection systems are limited in their forms compared with the sentences of a living language [1]. With these points in mind, this study attempts to establish a methodology for improving the capability of SIS-RT for traceability analysis.

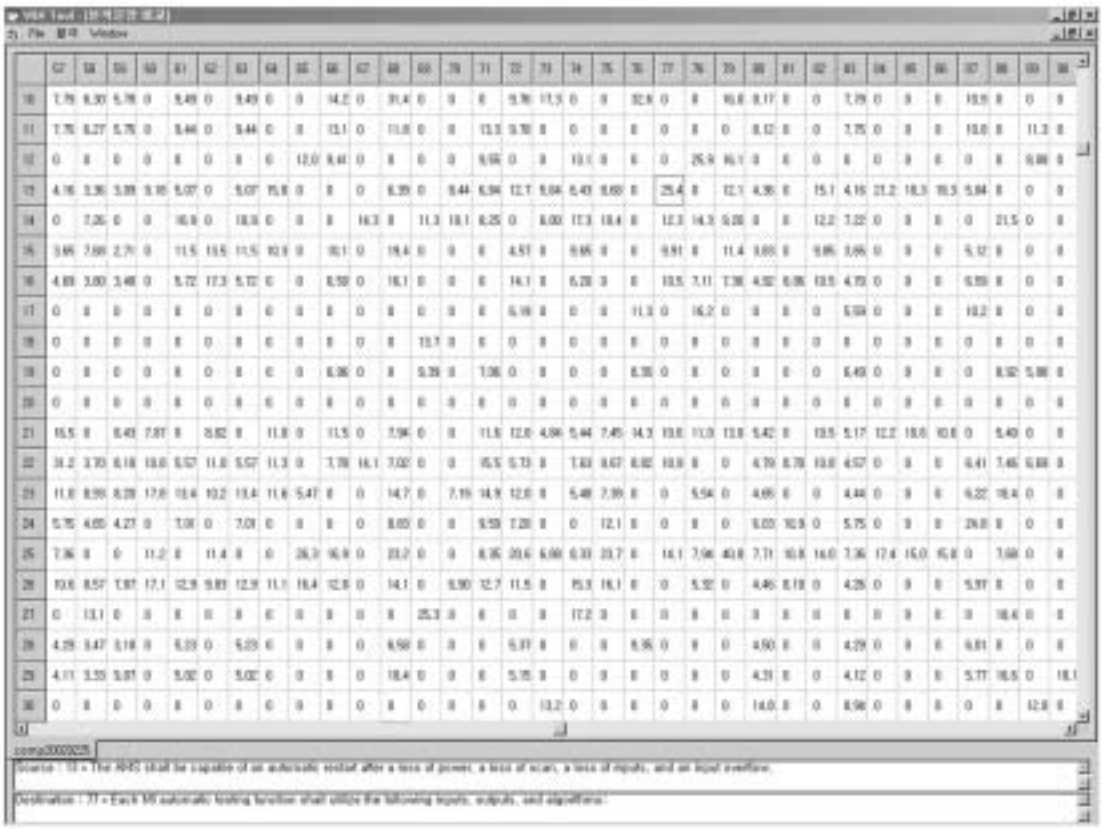

Fig. 2. Traceability Analyses Window of SIS-RT (improved)

2. Related Works on Natural Language Processing

2.1. Statistical Analysis - Cosine Vector Similarity Formula

In 1957, Luhn noted that an information retrieving system could be created by comparing specific words with the words within a query. Once certain important words (terms) are extracted via an analysis on the documents subject of a search, each document can be expressed in a vector form, according to the existence of the terms, as follows:

$$D = (t_1, t_2, t_3, \cdots, t_n) \tag{1}$$

Where, D is a term vector of a specific document, and

t_k is 0 or 1 whether or not the document contains specific terms (k = 1, 2, ⋯, n).

An information request or a query can also be expressed by a term vector, as follows:

$$Q = (q_1, q_2, q_3, \cdots, q_n) \tag{2}$$

Where, Q : is a term vector of a query, and

q_k : is 0 or 1 whether or not the query contains specific terms (k = 1, 2, ⋯, n).

The simplest method defining the similarity is by computing the value of similarity with the number of terms coexisting in a document and a query. In this case, the similarity is represented by the following formula:

$$Similarity(D,Q) = \sum_{k=1}^{n} t_k q_k \quad (3)$$

By assigning different weighting factors to each term, the measurement can be more effective than the mere use of 0s and 1s. Such weighting factors can be assigned using many different methods. The normalization of vectors is a commonly used method. By this procedure, the similarity between a document and a query are obtained from a cosine vector similarity formula, as follows:

$$Similarity(D,Q) = \frac{\sum_{k=1}^{n} w_{qk} w_{dk}}{\sum_{k=1}^{n} (w_{qk})^2 \cdot \sum_{k=1}^{n} (w_{dk})^2} \quad (4)$$

Where, w_{qk} : is a weighting factor of k-th term in the query, and

w_{dk} : is a weighting factor of k-th term in the document (k = 1, 2, ⋯, n)

2.2. Linguistic Analysis

Unlike English, the Korean language has free word order (scrambling) and inflections. In addition, there are many ellipses of essential sentence components. Furthermore, Korean has agglutinative characteristics wherein a word is formed with an essential morpheme and a formal morpheme [1].

Consequently, linguistic methods are more favorable than statistical methods for the analysis and the processing of Korean. Linguistic methods modify and simplify the grammars of natural languages and then analyze the language. Typical grammars are as follows [2]:

- *Phase structure grammar (Chomsky)*
- *Unification grammar*
- *Dependency grammar (Tesniere)*
- *Case grammar (Fillmore)*

Among these grammar forms, dependency grammar and case grammar are favored for their suitability to scrambling and omitting. There have been many studies on natural language processing systems using these grammars, particularly in the department of computer science at KAIST.

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

This study proposes a methodology for analyzing traceability based mainly on the concepts of case grammar. As the SIS-RT traceability analysis tool accepts sentences as inputs, the bases of the approach are comparisons between two sentences.

In the upgrading of the tool for English documents, similarity is computed based merely on the number of words in common for two compared sentences (see 2.1). In analyzing Korean, however, it is possible to grasp the cases of substantives with the information about the case frames of verbs and the postpositions added to the substantives [8]; therefore, substantives of the same cases can be compared to obtain the similarity of two sentences.

The procedure to obtain the similarity embodied in SIS-RT is as follows (Fig. 3-1):

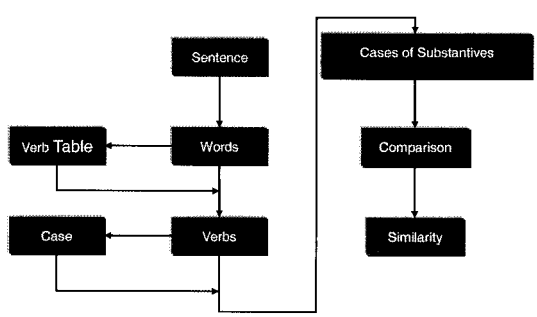

Fig. 1. Flow Chart of the Analysis

Table 1. Requisite Cases for Nuclear Field

Serial #	Requisite Cases	Definitions
1	Agent	The Starter of an event
2	Object	A substance which moves or changes A substance whose location or existence is considered
3	Instrument	An immediate cause of an event
4	Source	A start point of a moving substance
5	Goal	A destination of a moving substance
6	Locative	A spatial point where an event occur
7	Path	A channel through which a substance moves
8	Condition	A condition for the occurrence of an event

Table 2. Postpositions for Nuclear Field[26]

Serial #	Case Markers	Serial #	Case Markers
1	이, 가, 은, 는	6	로, 으로
2	을, 를	7	로부터, 에게서
3	에, 에서, 에서는	8	_시, _시에, _시에는
4	에, 에게, 에는, 에게는	9	을_통해, 를_통해
5	로, 로써, 으로써		

Table 3. The Verb Table

Group	Subgroup	Verbs
		Case frames
⋯	⋯	⋯
전송 (15.0)	15.1	전송한다, 전송하며, 전송하고, 전송하여, 전송하면
		[1 1][2 2][7 9][5 6]
	15.2	-
		-
	15.3	전송되어야, 전송된다
		[2 1][4 7][5 6][7 9]
⋯	⋯	⋯

- Distinction of the verbs using the verb table
- Discernment of deep cases with the information from the case frames and postpositions
- Expression of each sentence in a vector form
- Calculation of the similarity by comparing substantives of the same cases

▶ The Verb Table

The proposition core of a simple sentence is composed of one or more substantives and a predicate [5]. Thus, for the purpose of analyzing the case structure, predicates (verbs) must be extracted from the sentences.

In this study, 93 verbs were grouped into 32 groups according to their meaning. Each group was subcategorized into 3 subgroups, corresponding to modalities (Table 3-3).

Once the verb table is constructed, the verbs can be extracted from the input sentences.

▶ Postpositions

In the Korean language, a noun can be followed by an auxiliary verb, a suffix, or a particle [6]. As Korean postpositions are limited in number, they can be analyzed and processed easily. With a sufficient number of sentences, it is possible to classify the categories of case particles and the cases they represent (Table 3-2).

▶ Case Frames

A verb has one or more case frames (Fig. 3-2). 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 [4]. One verb can have different case frames,

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

Fig. 3. Verbal Case Frames [7]

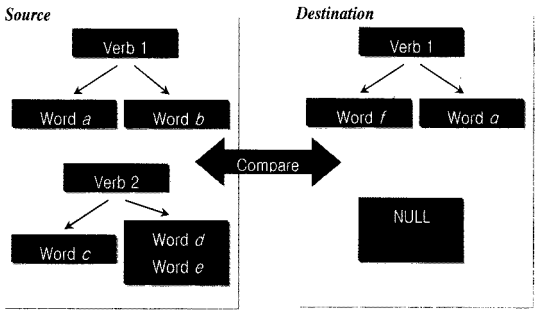

Fig. 3. A Comparison of Two Sentences Using Case Grammar

determined by the morphemes added to it, and there exist some rules. Thus, the analyses of case frames are simple.

Once the case frame of a verb is determined, the deep cases of substantives can be determined. Then, by comparing nouns of same cases in two sentences, a more detailed traceability analysis can be achieved.

Once a sentence is entered as an input, it is divided into terms. Each word is then compared with the verb table to distinguish a verb from the others. A case frame for the verb is brought for the verb. A case frame has the form of [case number, postposition number]. For example, if it was [7 3], the word that ends with a postposition of " 3 " will have the case number " 7. " In the same manner, every word will have its own case number (Fig. 3-3). Next, analyzed sentences are expressed in vector form in order to compute their similarity. The proposed method differs from the method of using the cosine vector similarity formula alone in that the proposed method does not only concern the existence of specific words while the existing method does, but it also concerns the semantic roles of each word. In addition, the elements at the same positions of the source and destination vectors have the same case numbers (semantic roles). Thus, terms having the same semantic roles can be compared.

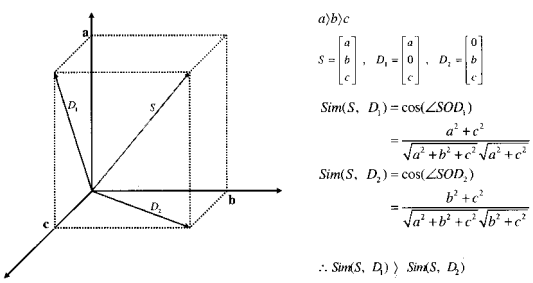

Fig. 3. Cosine Vector Similarity Formula with Term Weightings

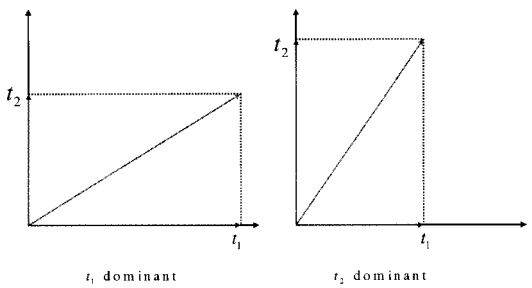

Fig. 5. Effect of Term Weightings

Once the two sentences are expressed in vector form, they are entered into the cosine vector similarity formula in order to compute the similarity value. The cosine vector similarity formula itself merely computes the closeness of two vectors using the inner product of vectors (Figs. 3-4, 3-5). If sentences are expressed in vector form, it is reasonable to use the cosine vector similarity formula.

4. Results and Discussion

Sentences were compared using the procedures described in the previous section. The following is an example:

Source:중성자속 출력이 프로그램 설정값 이상의 비율로 증가하거나 중성자속이 기 설정된 최대값에 도달하게 되면 원자로가 정지된다.

Destination: 가변 과출력 트립은 원자로 중성자속 출력이 프로그램 설정값 이상의 급격한 비율로 증가하거나 기 설정된 최대값에 도달하게 되면 원자로를 정지시키기 위하여 개시된다.

According to the case frames, the two sentences are restructured, as shown in Fig. 3-3. Then, the sentences can be expressed in vector form, as follows:

S = [01111101010111000011000]

D= [11111111111111111101111]

Next, we put the sentences into the cosine vector similarity formula to obtain the similarity value.

$$Similarity(S, D) = \frac{S \cdot D}{\|S\| \|D\|} = 69.3\%$$

More examples are classified into 4 groups.

▶ Sentences having different structures and similar meanings

S : 비교논리에는 발전소 기동 및 정지 시 불필요한 원자로 트립을 방지하기 위해 트립 기능을 차단시키는 운전우회 기능이 제공된다.

D : 운전우회는 발전소 기동 및 정지 시에 불필요한 발전소 트립을 방지하고, 저 출력 운전 시험을 허용하기 위한 기능이다. (38.7% → 46.3%)

▶ Sentences having similar structures and meanings

S : 각 채널의 해당 계열의 비교논리 프로세서 출력은 안전 데이터 링크를 통해 동일 채널 및 타 채널의 해당계열 동시논리 프로세서로 전달된다.

D : 비교논리 프로세서의 비교논리 출력은 광케이블을 사용한 안전 데이터 링크를 통해 4개 채널의 동시논리 프로세서로 전달된다. (59.7% → 99.4%)

▶ Sentences having different structures and meanings

S : 동시논리 프로세서는 동일 채널 및 타 채널의 해당계열 비교논리 프로세서로부터 전송된 4개의 트립 상태 신호와 동일 채널의 ATIP으로부터 전송된 트립채널우회(trip channel bypass) 신호를 이용하여 출력 신호를 결정한다.

D : 각 채널의 해당 계열의 비교논리 프로세서 출력은 안전 데이터 링크를 통해 동일 채널 및 타 채널의 해당계열 동시논리 프로세서로 전달된다. (28.7% → 0%)

▶ Sentences having similar structures and different meanings

S : 안전주입 작동신호는 가압기 저압력 혹은 격납용기 고압력 신호에 의해 발생되며

D : 격납용기 격리 작동신호는 격납용기 고압력 혹은 가압기 저압력 신호에 의해 발생되며 (92.8% → 70.7%)

5. Conclusions

This study proposed a method that uses a case grammar approach to improve the performance of the SIS-RT tool. The main advantage of the proposed method is that analyzed documents are specified in the nuclear domain, making the rules required for an analyses relatively simple, resulting in a practical, high performance system. The proposed method showed satisfactory results for some initial cases. It is remarkable that the proposed method can obtain a higher similarity than the statistical method, for a case involving different structures and similar meanings. However, if two sentences have similar structures and are composed mainly of the same words, as in the case of similar structures and different meanings, the proposed method was not found to be superior to the statistical method.

This originates from the nature of proposed method, which essentially analyzes the structure of sentences. Nevertheless, the proposed method can be considered a more semantic-oriented method than the statistical method, because it restructures sentences based on the semantic roles of words or phrases in the analyzed 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).