Detection of Synonyms in Specification Documents

Yasushi Kawai*, Tomohiro Yoshikawa*, Takeshi Furuhashi*, Eiji Hirao†, Ayako Kuno† and Tomohisa Gotoh†
*Nagoya University, Japan
Furo-cho, Chikusa-ku, Nagoya 464-8601 JAPAN
kawai@cmplx.cse.nagoya-u.ac.jp, yoshikawa@cse.nagoya-u.ac.jp, furuhashi@cse.nagoya-u.ac.jp
†NEC corporation
1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa 211-8666, JAPAN
e-hirao@bu.jp.nec.com, a-kuno@ah.jp.nec.com, t-gotoh@bq.jp.nec.com

***Abstract*—Recently, the document information managed in companies becomes complex and various more and more. Specification documents are used for the technical transfer and inheritance of manufactures and services. However, the description and the meaning of the component words in specifications are often inconsistent or multiple, because a specification document is made by the persons in charge of various parts. Then the readers may misunderstand the contents of specifications by its inconsistency. This paper focuses on synonyms, multiple description of words for a meaning or a word, in specification documents and proposes an extraction method of them considering the co-occurrence words of each morpheme in compound words. This paper applies the proposed method to a test data, in which some words in an actual specification document are replaced with words, and studies the effectiveness of the proposed method.**

I. Introduction

Recently, the document information managed in companies becomes complex and various more and more with the development of the computer performance and the accumulation of technologies. Specification documents are used in various situations for the technical transfer and inheritance of manufactures and services. Specifications are the documents describing the flow of operations and each process, which include the procedures of the project and the arrangement based on the agreements, and they are necessary when people want to transfer the complicated contents of work or the requirements to other people in charge or customers. However, a specification document is usually made by plural persons in charge of various parts, which often causes multiple description or definition of the component words. The readers may misunderstand the contents of specifications by these multiple-description/definition. This paper focuses on synonyms in specifications, which are the multiple description of words for the same meaning, and tries to extract the candidates of them in the document. Showing the candidates of synonymous will support and reduce the burden of reviewing the specifications greatly. A lot of research has been reported for the detection of the synonyms in documents [1], [2]. One of these approaches is to use dictionaries such as thesaurus. However, the synonyms in specification documents are those just in the small corpus, i.e., the target document, and a lot of unknown words or technical terms which thesauri do not have can be the synonyms. Another approach is based on the distributional hypothesis [3], [4] that synonyms have similar contextual information. Based on this hypothesis, the contextual information for each word is represented as the co-occurrence vector, which is the co-occurrence relation with other words in the document, and the degree of similarity between words are quantified based on the index such as the cosine similarity of the co-occurrence vectors. However, it is difficult to quantify the similarity of the contextual information effectively because the cosine similarity calculates the exact matching of words without considering the similarity of meaning or use.

This paper proposes an extraction method of synonyms in specification documents, which are difficult to be extracted based on thesauri. The proposed method employs the similarity of the contextual information described above, then it is expected that the specific synonyms which are shown just in the document can be extracted. The technical terms, most of which are compound words, are intended in the proposed method by dividing the compound words into the morphemes which are the smallest semantic units in a language. Moreover, the proposed method employs the aggregation method of the words in the co-occurrence vector based on the thesaurus to consider the similarity of meaning or use between the co-occurred words. The compound words are also aggregated using the dominant weight of each morpheme.

This paper applies the proposed method to an actual specification, in which half of some words are replaced into words to put artificial synonyms in the document, and studies the effectiveness of the proposed method by the extraction performance of the candidates of synonyms for the replaced words.

II. Related Work

A lot of research to extract synonyms in documents automatically has been reported. Terada et al. [5] detected synonyms aiming for the correspondence between an abbreviated word and its basic form, using the local information around the words as the contextual information. Wang et al. [6] studied the validity of the method which gave the different weights to four features of the contextual information, the adjective modifying a noun, the verb taking a noun as its object/subject, and the neighboring words of a noun. They also proposed the re-rank method of the calculated word pairs using the word-similarity network. Though the aim of these studies is

similar to this paper in terms of extracting synonyms, the target documents are basically general in these studies while those are specification documents with a lot of synonyms of technical terms in this paper. Terada et al. [7] intended to extract technical terms and their synonyms in a particular area using queries and proposed the tool to show the candidates of synonyms. This method is the extension of [5] to Japanese with some devices in the use of the contextual information. It is different from this paper in terms of dealing with a compound word as one word.

III. Proposed Method

This section shows the flow of extracting synonyms by the proposed method.

A. Generation of Co-occurrence Vector

First, this method generates co-occurrence vectors to all nouns in the document which are the target of synonyms to be extracted. As the contextual information, this method uses the words which appear in the same sentence except for itself. The elements of the vector are all nouns, verbs and adjectives which appear at least once in the document, and the value of each element is the number of occurrences. When successive nouns appear, it regards them as a compound word and one element of the vector.

B. Calculation of Dominant Weight

This sub-section calculates the dominant weight of each morpheme in the compound words. A compound word is divided into morphemes and the dominant weight for each morpheme is given by the strength of the word. This strength of a word means the uniformity of the meaning of compound words when the word, i.e., the morpheme, becomes a part of compound words, and the uniformity is quantified by the co-occurrence of the compound words. The hypothesis here is "The more strongly a word dominates, the more similarly the compound words are used." TABLE I shows the co-occurrence vector of "system" and "change" which are the morphemes of the compound word "system change." The dominant weight is calculated as shown in eq.(1).

$$DominantWeight = \frac{N_x}{N_y} \tag{1}$$

In eq.(1), N_y is the number of co-occurrence words which appear one or more times in the co-occurrence vectors of the compound words, e.g. 8 in TABLE I (a), and N_x is the number of co-occurrence words which appear more than once in the vectors, e.g. 2, "proposal" and "update," in TABLE I (a). When dominant weight becomes "1," it means no matter which words are combined with, the use of the compound words, i.e., the co-occurred words, is similar. In TABLE I, the dominant weight of "system" is $2/8 = 0.25$, and that of "change" is $5/6 \simeq 0.83$. The dominant rate is also calculated here, which becomes as follows;

$$\begin{aligned} &DominantRate(\text{“}system\ change\text{”}): \\ &\text{“}system\text{”} : \text{“}change\text{”} = \frac{0.25}{0.25+0.83} : \frac{0.83}{0.25+0.83} \\ &\qquad = 0.23 : 0.77 \end{aligned} \tag{2}$$

TABLE I
Example of Co-occurrence Vectors of Morphemes

(a) "system"

	occur	proposal	use	update	base	separate	network	user
system error	1	0	0	0	0	0	0	0
operating system	0	0	0	1	1	0	1	0
information system	0	1	0	0	0	1	0	0
system change	0	1	0	0	0	0	0	0
system tool	0	0	1	1	0	0	0	1

(b) "change"

	control	follow	setting	proposal	considering	operation
contents change	1	1	0	1	1	1
system change	0	0	0	1	0	0
addition and change	0	1	0	0	0	1
prohibition term of change	0	0	1	1	0	0
organization change	1	1	0	0	0	1
demand of change	0	0	0	0	1	1

C. Generation of Concept Vector

This sub-section generates concept vectors using dominant rate described in III-B. This study defines "concept" which is the aggregated group of words in the vectors based on a thesaurus. This paper uses "Nihongo Dai-Thesaurus" [8] as the thesaurus. This thesaurus has three hierarchical groups of words. This paper calls the highest level group as "large concept," middle one as "middle concept," and lowest one as "small concept." The small concept is the minimum unit of words to be aggregated, and the small concepts and the words belonging to them are aggregated to the upper concept, i.e., middle concept. Compound words are considered as one element of co-occurrence vectors, and aggregated to the concepts considering dominant rate. Fig.1 shows an example of the aggregation of the compound word "system change" to the concepts based on the dominant rates. In this example, "system change" in III-B is aggregated to the concepts that "system" belongs to as 0.23, and to those "change" belongs as 0.77. "system" belongs to "computers," "information science," and so on in the small concept, and "change" does to "reform." When a word co-occurred with "system change" three times, the values of the element in the concept vector for this word become $3 \times 0.23 = 0.69$ for "computers," "information science" and $3 \times 0.77 = 2.31$ for "reform," respectively.

Co-occurrence vector

…	…	system change	…	…
…	…	3	…	…

Concept vector

…	computers	information science	reform	…
…	0.69	2.31	2.31	…

Fig. 1. Example of Aggregation to Concepts

D. Calculation of Similarity

Equation (3) shows the modified cosine similarity to calculate the similarity between concept vectors. In eq.(3), k is the adjustment term for the consideration of low frequency words, because it is often seen that low frequency pairs of words tend to have high similarity.

$$\frac{\vec{x} \cdot \vec{y}}{\sqrt{|\vec{x}|^2 + k^2}\sqrt{|\vec{y}|^2 + k^2}} \tag{3}$$

E. Filtering of Candidates of Synonyms

This method can filter the candidates of synonymous by appending some rules arbitrarily. The rules which users can select are shown below.

(a)Pairs which do not appear in a common sentence
(b)Words which do not contain a number
(c)Pairs of words whose detailed classification of a part of speech are matched
(d)Pairs of words whose concept or one of the concepts of the morphemes are matched
(e)Using string matching or edit distance as the threshold

(a) is the rule to exclude the pairs of words which appear in a common sentence from the candidates because the co-occurrence of these pairs tends to be similar. Besides, synonyms rarely appear in a common sentence except for the explanation of the abbreviation. In (b), the words containing a number can be hardly synonyms, and there are a lot of sentences that only the number part is different in specification documents, which makes the similarity of those words high. (c) verifies whether detailed classification of a part of speech, e.g., proper nouns, common nous and concrete nous, are matched. (d) and (e) are the filtering by the meaning or the description.

IV. Experiment

A. Specification Document and Synonyms

The proposed method was applied to "Outsourcing operations specification of information center system for cancer measures in 2009 (in Japanese)" [9] from the Ministry of Health, Labour and Welfare (Japan). We put artificial synonyms in the document by replacing half of the words in the left column of TABLE II into other words shown in the right column of TABLE II. In this experiment, the performance of the extraction of the candidates of synonyms is evaluated by the rank order of those words for the candidates in the all pairs of nouns in the document. The synonyms in TABLE II are originally written in Japanese. A and B in TABLE II are high frequency words, C and D are middle frequency words, and E and F are low frequency words.

TABLE II
CORRECT WORDS MADE BY REPLACEMENT

pair	before replacement		after replacement	
	word	number of occurrences	word	number of occurrences
A	management	89	supervision	75
B	operation trustee	61	management trustee	56
C	making	16	establish	23
D	obstacle	17	abuse	16
E	probe	8	consideration	6
F	keeping space	10	preserving space	5

B. Effects of Aggregation by Concepts

First, the effect of the aggregation by the concepts was studied comparing with no aggregation, i.e., using the co-occurrence vectors described in III-A. Here, k in eq.(3) was set to 0, and the dominant weight was not taken into consideration and the weight of each morpheme in compound words was equivalent. TABLE III shows the similarity which is calculated by eq.(3) and the rank order of the artificial synonyms in TABLE II. The sorting of the rank is given by the descending order of the similarity. "words" in TABLE III means no aggregation, and "large concepts" means that large concepts were used for the aggregation in the concept vector described in III-C. The upper value in TABLE III is the value of similarity for each pair, and the lower value is the rank order in all candidates. The number of pairs in the document was 2,593,503, which means the pair of correct synonyms A, "management" and "supervision," appear as the 2792th candidate of synonyms in 2,593,503 pairs in "words." In TABLE III, the rank orders using the concept vector are superior to those using the co-occurrence vector in all synonyms. It is important to raise not only the similarity but also the rank order, because the reviewer looks over the candidates of synonyms from the top in rank. More over, we can see that larger concepts show better result. One of the reason is that the co-occurrence vectors were too sparse to extract the proper contextual information because of the large amount of pairs. In the following experiment, this paper uses the large concepts for the aggregation whose result was the best in TABLE III.

C. Study on Consideration of Dominant Weight

Second, the effect of the dominant weight was studied with the following conditions.

Method 1: $k = 0$ in eq.(3)
Method 2: $k = 3$ in eq.(3)
Method 3: $k = 3$ in eq.(3), and filtering rule (a), (b) and (c) in III-E were applied.

TABLE III
EFFECT OF AGGREGATION BY CONCEPTS

pair	words	small concepts	middle concepts	large concepts
A	0.914	0.960	0.968	0.982
	2792	1508	1382	1569
B	0.905	0.964	0.968	0.991
	2914	1433	1394	1035
C	0.488	0.947	0.948	0.975
	106282	1871	1985	2125
D	0.780	0.921	0.926	0.970
	8359	2800	2659	2035
E	0.761	0.919	0.926	0.967
	7338	3196	2989	2860
F	0.690	0.866	0.885	0.950
	13144	11092	7367	6348

TABLE IV shows the result of the similarity and rank order in Method 1 to Method 3 without consideration of the dominant weight. The number of pairs in Method 1 and 2 was 2,593,503, and that in Method 3 was 618,297. In TABLE IV, the rank order of Method 2 and 3 were higher than that of Method 1 in all synonyms. The adjustment term k and the filtering rules in the proposed method worked very well especially in the high frequency words. The exclusion of the noise candidates, high similarity because of low frequency, appearing in the same sentence, and just the difference of the number part, made the rank order of the correct synonyms higher.

TABLE IV
RESULT WITHOUT CONSIDERATION OF DOMINANT WEIGHT

pair	Method 1	Method 2	Method 3
A	0.982	0.982	0.982
	1569	174	31
B	0.991	0.990	0.990
	1035	12	2
C	0.975	0.974	0.974
	2125	444	14
D	0.970	0.968	0.968
	2616	1003	46
E	0.967	0.961	0.961
	2860	1820	155
F	0.950	0.937	0.937
	6348	7078	1166

TABLE V shows the result using the dominant weight described in III-B. The value after the rank order is the difference from TABLE IV. Most of the cases in TABLE V show better result. The consideration of dominant weight for compound words worked well especially in the low frequency words because the proper contextual information could be used.

TABLE V
RESULT USING DOMINANT WEIGHT

pair	Method 1	Method 2	Method 3
A	0.982	0.982	0.982
	1504 (-65)	164 (-10)	3 (0)
B	0.990	0.990	0.990
	981 (-54)	7 (-5)	2 (0)
C	0.974	0.974	0.974
	2038 (-87)	449 (5)	14 (0)
D	0.967	0.966	0.966
	2694 (78)	1096 (93)	51 (5)
E	0.970	0.964	0.964
	2422 (-438)	1344 (-488)	81 (-74)
F	0.955	0.942	0.942
	4451 (-1897)	4699 (-2379)	674 (-492)

V. CONCLUSION

This paper proposed an extraction method of synonyms in specification documents, which were difficult to be extracted based on thesauri. The dominant weight to consider the meaning of compound words, the aggregation of words in the co-occurrence vector based on the concept, the adjustment term to consider low frequency words, and filtering rules to exclude noise candidates were introduced in this paper. The proposed method was applied to an actual specification which had artificial synonyms by replacing words, and the results showed the proposed method worked very well to extract synonyms in the specification documents. For the further work, the study on other calculation methods for dominant weight, the proper value of the adjustment term k, and more effective exclusion of noise candidates will be needed. We will also apply the proposed method to other specification documents.

REFERENCES

[1] Dekang Lin,"Automatic retrieval and clustering of similar work" Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational linguistics (COLING-ACL'98):786-774, 1998.

[2] P. D. Turney,"Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL" In Procs. of the Twelfth European Conference on Machine Learning (ECML-2001), pages 491-502, Freiburg, Germany, 2001.

[3] Harris Z,"Distributional structure" Word 10 (23): 146-162, 1954.

[4] Lee, L., "Measures of distributional similarity," Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 25-32, 1999.

[5] Akira Terada, Takenobu Tokunaga, Hozumi Tanaka, H,"Automatic expansion of abbreviations by using context and character information" Inf. Process. Manage., 40(1),pp. 31-45, 2004.

[6] Yuxin Wang, Nobuyuki Shimizu, Minoru Yoshida, Hiroshi Nakagawa,"Automatic Synonym Acquisition through Word Similarity Network (in Japanese)," Technical Report of Natural Language Processing 2008(46), 7-14, 2008.

[7] Akira Terada, Minoru Yoshida, Hiroshi Nakagawa,"A System for Constructing a Synonym Dictionar (in Japanese)," Journal of Natural Language Processing 15(2), 2008-04, pp. 39-58, 2008.

[8] T. Yamaguchi, Nihongo Dai-Thesaurus CD-ROM (in Japanese), Tokyo, Taishukan Shoten, 2006.

[9] Ministry of Health, Labour and Welfare : "Outsourcing operations specification of information center system for cancer measures in 2009 (in Japanese)," http://www.mhlw.go.jp/sinsei/chotatu/chotatu/kankeibunsho/090123/