Automatic Synonym Acquisition Based on Matching of Definition Sentences in Multiple Dictionaries

Masaki Murata[1], Toshiyuki Kanamaru[2], and Hitoshi Isahara[1]

[1] National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{murata, isahara}@nict.go.jp
http://www.nict.go.jp/jt/a132/members/murata/

[2] Kyoto University, Yoshida-nihonmatsu-cho,
Sakyo-ku, Kyoto, 606-8501, Japan
kanamaru@hi.h.kyoto-u.ac.jp

Abstract. Studies on paraphrasing are important with respect to various research topics such as sentence generation, summarization, and question-answering. We consider the automatic extraction of synonyms (which are a kind of paraphrase) through the matching of word definitions from two dictionaries, and describe a new method for extracting paraphrases. Higher precision was obtained than with a conventional frequency-based method. The new method provided a precision rate of 0.764 for the top 500 data pairs and 0.220 for 500 randomly extracted data pairs when only synonyms were considered a correct answer. It provided a precision rate of 0.974 for the top 500 data pairs and 0.722 for 500 randomly extracted data pairs when hypernyms and similar expressions were also considered correct answers. Our method should be useful for other studies on paraphrase extraction.

1 Introduction

Studies on paraphrasing [6, 2] have had important consequences in various domains such as sentence generation, summarization, and question-answering [3, 12]. Likewise, studies on paraphrase extraction are also important. In this paper, we discuss the automatic extraction of synonym expressions which can be considered a kind of paraphrase. We extract synonym expressions by matching definitions of the same word from two dictionaries. In this work, we studied the extraction of synonym expressions in the Japanese language.

For example, we examined the definition sentences for the word *abekobe* meaning "reverse". Two Japanese dictionaries gave the definitions shown in Figure 1 for the word. We expected to extract pairs of expressions having the same meaning when we compared the two definitions, since they both defined the same word and thus had the same meaning. We compared the two definition sentences and obtained the results shown in the figure. From the results, we deter-

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 293–304, 2005.

Definition of "reverse" in Dictionary A:

junjo	,	*ichi*	*nado*	*-no*	*kankei*	*-ga*	*sakasama-ni*	*irekawat*
(order)	(,)	(location)	(etc.)	(of)	(relation)	nom	(upside-down)	(change places)
-teiru								
(-ing)								

(The relationship of the order, location, and so on is changed upside-down.)

Definition of "reverse" in Dictionary B:

junjo	,	*ichi*	,	*kankei*	*-ga*	*hikkuri-kaet*	*-teiru*
(order)	(,)	(location)	(,)	(relation)	nom	(be overturned)	(-ing)

(The relationship of the order and location is overturned.)

Results of comparing the two definitions

junjo , *ichi*	*nado*	*-no*	*kankei -ga*	*sakasama-ni*	*irekawat*	*-teiru*
	(etc.)	(of)		(upside-down)	(be changed)	
	,			*hikkuri-kaet*		
	,			(be overturned)		

Fig. 1. Example of rule extraction for paraphrasing

mined that *nado-no* "etc." and "," were interchangeable, as well as *sakasama-ni irekawatte* "be changed upside-down" and *hikkuri-kaet* "be overturned". In short, our method for extracting synonym expressions is to extract synonym expressions by matching definition sentences from two dictionaries having the same content.

The advantages of our method can be summarized as follows.

- Although synonym expressions were extracted from text pairs having the same content in previous studies, there have been no studies where definition sentences in multiple dictionaries were considered text pairs having the same content and synonym expressions were extracted from them. This paper is useful in showing that many synonym expressions can be extracted from definition sentences in multiple dictionaries.
- In this paper, we propose a new method, which is useful for extracting synonym expressions. We show, based on our experiments, that this method is more effective than several comparable methods. This method can also be used for other studies on the extraction of synonym expressions.

2 Method of Extracting Synonym Expressions Based on Matching Two Dictionaries

In this study, we extracted synonym expressions by matching definitions of the same word from two dictionaries: the Iwanami Japanese dictionary and the Daijirin Japanese dictionary.

We first aligned definition sentences for the same word that were extracted from the two dictionaries. When a word had more than one definition sentence,

Table 1. Examples of the results from definition sentence matching

Degree of matching	Word	Compared definition sentences
0.69	*appuappu* (gasp for breathe)	*mizu-ni obore-kakete, mogaiteiru sama* (to struggle by reason of drowning) *mizu-ni obore-kakete kurushimu sama* (to suffer by reason of drowning)
0.20	akarasama (frank)	*kyu-na sama* (the state of suddenness) *tsutsumi-kakusanaide, hakkiri arawasu sama* (the state of expressing something plainly without concealing one's feelings)

we assumed one-to-one alignment and aligned the pair of definition sentences having the best degree of matching.

We separated each definition sentence into several words by using the JUMAN Japanese morphological analyzer[4] and arranged each word on each line. We detected matching parts and non-matching parts by using the UNIX diff command [8, 5, 9]. We defined the degree of matching as

$$\text{Degree of matching} = \frac{N_{match} \times 2}{N_{all}}, \tag{1}$$

where N_{match} is the number of characters in the matching part and N_{all} is the total number of characters in the two definition sentences. The degree of matching takes a value from "0" to "1" with the value being larger when the matching part is larger.

When we performed the above alignment and matching of definition sentences, we obtained 57,643 definition sentence pairs. Some examples of the results from definition sentence matching are shown in Table 1. In the table, parts that differ are underlined and these were extracted as candidate synonym pairs.

We found some good synonym pairs such as the pair of *mogaiteiru* "struggle" and *kurushimu* "suffer", but also found some pairs that were not synonyms such as the pair of *kyuu-na* "suddenness" and *tsutsumi-kakusanaide, hakkiri arawasu* "expressing something plainly without concealing one's feelings". These results were not particularly accurate and could not be used as synonyms as they were.

Therefore, we next extracted better synonym pairs from candidate synonyms. We based this extraction on the following characteristics.

– Differing parts that are surrounded by lower-frequency words are better synonym pairs.
– Differing parts that occur more frequently are better synonym pairs.

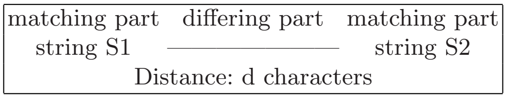

Fig. 2. Occurrence of differences

First, we considered the first characteristic, "differing parts that are surrounded by lower-frequency words are better synonym pairs." We assumed that a differing part would be surrounded by strings, $S1$ and $S2$, which were matching parts and the distance between $S1$ and $S2$ would be d characters[3] as shown in Figure 2. The probability, $P(S1)$ or $P(S2)$, of the occurrence of $S1$ or $S2$, respectively, in an inner region consisting of no more than d characters from $S2$ or $S1$ is approximately expressed as

$$P(S1) \simeq (d+1) \times \frac{Freq(S1)}{N} \tag{2}$$

$$P(S2) \simeq (d+1) \times \frac{Freq(S2)}{N}, \tag{3}$$

where $Freq(S1)$ and $Freq(S2)$ are the respective numbers of occurrences for strings $S1$ and $S2$, and N is the total number of characters in the database. If we assume the probability $P(dfp, S1, S2)$ that the differing part (dfp) is good equals the probability that strings $S1$ and $S2$ do not appear in the situation shown in Figure 2, $P(dfp, S1, S2)$ can be expressed as

$$P(dfp, S1, S2) \simeq (1 - P(S1))(1 - P(S2)), \tag{4}$$

where we assume that $S1$ and $S2$ are independent of each other.

Next, we considered the second characteristic, "differing parts that occur more frequently are better synonym pairs." We have only to combine the probabilities in multiple situations. We assumed that when at least one of the multiple situations was correct, we would extract the differing part as a correct one. Since the differing part being correct is the complement of the case where all the situations for the differing part are incorrect, the probability $P(dfp)$ that the differing part (dfp) is correct is expressed as

$$P(dfp) \simeq 1 - \prod_{S1,S2} (1 - P(dfp, S1, S2)), \tag{5}$$

where we assume that each situation for the differing parts is independent of one another.

The extraction of synonym pairs is done by sorting candidate differences according to the value of the above equation and extracting the one having a higher value. In this paper, we refer to this method as *our method*.

[3] In this study, we used a longer length of characters in the differences as d.

3 Comparison Method

In this section, we describe the comparison method used in the experiments to evaluate the effectiveness of our method.

- Frequency method
  Extracted differing parts are sorted by their respective frequencies. A difference having a higher frequency is judged to be a more plausible synonym pair.
- Katoh's method
  It is based on Katoh et al's study [3]. Differences that satisfy the following conditional equation are first extracted and these are then judged to be plausible synonym pairs based on the frequency method.

$$\frac{N_{S1} + N_{S2}}{d} > 1, \tag{6}$$

  where N_{S1} and N_{S2} are the numbers of characters of $S1$ and $S2$, respectively. When the summation of the lengths of $S1$ and $S2$ exceeds the length of the differing part, we judge that the difference is not an accidentally extracted one and represents a plausible synonym pair.
- Combined method
  This is a combination of Katoh's method and our method. Differences that satisfy Equation 6 are extracted and then are judged based on Equation 5 as to whether they are plausible synonym pairs.

4 Experiments

We used our method to obtain synonym pairs. Examples of extracted differences are shown in Table 2, and examples of good extracted synonym pairs are shown

Table 2. Examples of extracting differences

-log(1-P)	Frequency	Preceding contexts	Differing parts		Succeeding contexts
4975	786	*shinpai ga naku* (without trouble)		, (,)	*nonbiri shiteiru* (peaceful)
2266	301	*dankai* (grade)	*ga* (is)	*no* (is)	*hikui koto* (low)
1528	234	*kinzoku* (metal)		*no* (-lic)	*gen* (string)
208.8	60	*inkoku ni* (knife)	*tsukau* (used)	*mochiiru* (utilized)	*kogatana* (for linocut)
162.6	22	*tadashii kaitou* (true answer)	*matawa* (or)	*ya* (or)	*kaishaku* (interpretation)
105.8	22	*seizou* (method)	*suru* (performing)	*no* (of)	*houhou* (production)

Table 3. Examples of extracted synonym pairs

tsutsu (while)	*nagara* (while)
honyuu doubutsu (a mammal)	*honyuu rui* (the mammals)
tyuuto (halfway)	*totyuu* (on the way)
gyou (job, work)	*shoku* (job, work)
naru (become)	*kawaru* (change)
hedatari (gap)	*sa* (difference)
tsuku (get to)	*tyoutyaku suru* (arrive)
de tsukutta (made by)	*no* (of)
kachiku (a domestic animal)	*gyuuba nado* (horses and cows etc.)
ga umai (be good at)	*ni takumina* (be skillful at)
daiji ni (important)	*taisetsu ni* (precious)
tsutaeru (tell, report)	*dentatsu suru* (tell, report)
tameni (for)	*mokuteki de* (for the purpose of)
hazurete iru (be out of)	*awanai* (do not match)
ku (eat)	*taberu* (eat, have)
genshou suru (decrease)	*sukunaku naru* (become fewer)

in Table 3. We were able to extract many word-level synonym pairs and phrase-level synonym pairs such as *ga umai* "be good at" and *ni takumina* "be skillful at". We could also extract some functional word pairs such as *tsutsu* and *nagara*, which have the same meaning of "while".

Next, we compared our method to comparison methods. These results are shown in Tables 4 and 6. Here, we judged as correct the extracted pairs (differences) that have a context where they are judged to be synonym pairs. Table 4 shows the precision for the top X pairs for each method. Table 6 shows the precision and the number of extracted synonym pairs.

"Precision" in Table 6 means the precision for 500 randomly extracted pairs. The "Number of extracted pairs" is the total number of extracted pairs. The "Expected number of extracted synonym pairs" was obtained by multiplying

Table 4. Precision (Top 500 pairs)

	Our method	Frequency method	Katoh's method	Combined method
Top 50	0.900 (45/ 50)	0.580 (29/ 50)	0.680 (34/ 50)	0.900 (45/ 50)
Top 100	0.870 (87/100)	0.560 (56/100)	0.620 (62/100)	0.870 (87/100)
Top 200	0.820 (164/200)	0.580 (116/200)	0.645 (129/200)	0.825 (165/200)
Top 300	0.790 (237/300)	0.583 (175/300)	0.657 (197/300)	0.780 (234/300)
Top 400	0.767 (307/400)	0.590 (236/400)	0.642 (257/400)	0.767 (307/400)
Top 500	0.764 (382/500)	0.588 (294/500)	0.616 (308/500)	0.738 (369/500)

Table 5. Precision (Top 500 pairs excluding cases where a difference on one side is a null expression)

	Our method	Frequency method	Katoh's method	Combined method
Top 50	0.960 (48/ 50)	0.880 (44/ 50)	0.920 (46/ 50)	0.980 (49/ 50)
Top 100	0.960 (96/100)	0.900 (90/100)	0.930 (93/100)	0.960 (96/100)
Top 200	0.950 (190/200)	0.910 (182/200)	0.905 (181/200)	0.950 (190/200)
Top 300	0.933 (280/300)	0.903 (271/300)	0.907 (272/300)	0.927 (278/300)
Top 400	0.917 (367/400)	0.907 (363/400)	0.895 (358/400)	0.907 (363/400)
Top 500	0.904 (452/500)	0.910 (455/500)	0.878 (439/500)	0.876 (438/500)

Table 6. Precision and number of extracted synonyms

	Our method	Katoh's method
Precision	0.220 (110/500)	0.400 (200/500)
Number of extracted pairs	67851	17104
Expected number of extracted synonym pairs	14927	6841

Table 7. Occurrence rates for several relationships

	Synonym	Hypernym	Similar expression	No relation
random	0.220 (110/500)	0.454 (227/500)	0.048 (24/500)	0.278 (139/500)
Top 50	0.900 (45/ 50)	0.100 (5/ 50)	0.000 (0/ 50)	0.000 (0/ 50)
Top 100	0.870 (87/100)	0.120 (12/100)	0.000 (0/100)	0.010 (1/100)
Top 200	0.820 (164/200)	0.165 (33/200)	0.000 (0/200)	0.015 (3/200)
Top 300	0.790 (237/300)	0.190 (57/300)	0.000 (0/300)	0.020 (6/300)
Top 400	0.767 (307/400)	0.212 (85/400)	0.003 (1/400)	0.018 (7/400)
Top 500	0.764 (382/500)	0.206 (103/500)	0.004 (2/500)	0.026 (13/500)

"Precision" and "Number of extracted pairs", and is the expected number of synonym pairs that each method will be able to extract. Since Katoh's method uses elimination by Equation 6, the total number of extracted pairs in the method is smaller than in our method. Since the frequency method does not use elimination by Equation 6, the total number of extracted pairs is the same as in our method.

Table 8. Occurrence rates for several relationships (excluding cases where a difference on one side is a null expression)

	Synonym	Hypernym	Similar expression	No relation
random	0.313 (106/339)	0.274 (93/339)	0.071 (24/339)	0.342 (116/339)
Top 50	0.960 (48/ 50)	0.020 (1/ 50)	0.000 (0/ 50)	0.020 (1/ 50)
Top 100	0.960 (96/100)	0.010 (1/100)	0.000 (0/100)	0.030 (3/100)
Top 200	0.950 (190/200)	0.030 (6/200)	0.000 (0/200)	0.020 (4/200)
Top 300	0.933 (280/300)	0.033 (10/300)	0.007 (2/300)	0.027 (8/300)
Top 309	0.932 (288/309)	0.032 (10/309)	0.006 (2/309)	0.029 (9/309)

As shown in Table 4, the precision of our method and that of the combined method using Equation 5 were higher than with the other methods. We thus found that our proposed Equation 5 was effective.

Comparing the frequency method and Katoh's method, we found that Katoh's method provided higher precision. Thus, the deletion of candidates through Equation 6 in Katoh's method was effective.

Table 5 shows the results when we excluded cases where a difference on one side was a null expression. When a difference on one side is a null expression, the differences are not likely to be synonyms. Therefore, the results in Table 5 were better than those in Table 4.

In Table 6, the precision when 500 pairs were randomly extracted was 0.22 with our method and 0.40 with Katoh's method. This was because Katoh's method deleted unreliable candidate pairs through Equation 6. However, fewer synonyms were extracted with Katoh's method than with our method. Katoh's method has a shortcoming in that it fails to extract many synonyms. The estimated number of extracted synonyms with our method was 15000, and the precisions were 0.764 for the top 500 data (Table 4) and 0.220 for the 500 data that were extracted randomly (Table 6).

We next performed a more detailed examination using the results extracted by our method, which provided good results in the above experiments. The results are shown in Table 7. In the examinations, we counted the number of pairs having contexts that enabled them to be judged as either synonym pairs, hypernym pairs, similar to each other, or having no relationship. In the table, "random" indicates the results for 500 randomly extracted pairs and "Top X" indicates the results for the top X pairs. Table 8 shows the results from Table 7 when cases where a difference on one side was a null expression were excluded. Examples of expressions that were judged to be hypernyms or similar expressions are shown in Tables 9 and 10, respectively. This examination was done because most of the errors that were not judged to be synonyms were hypernyms or similar expressions.

When we considered the pairs that were hypernyms or similar expressions, as well as synonyms, to be correct, the precision rates became very high. For all the data, the precision rose to 0.722 (= 1 − 0.278). When we excluded cases where a difference on one side was a null expression, the precision became 0.658

Table 9. Examples of pairs having a hypernym relationship (a left expression meaningfully includes a right expression)

Preceding contexts	Differing parts		Succeeding contexts
me ya kuchi (open eyes and mouth)	*nado* (etc.)		*wo kyuuni hiraku sama* (suddenly)
nihiki	*ijou* (more than)		*no kaiko ga* (two silkworms)
houkou wo shimesu		*dai* (broad)	*houshin* (policy indicating directions)
takakkei no (polygon's)	*subete no* (all)	*kaku* (each)	*tyouten ga* (the vertex/vertexes)
hana ga (flowers)	*zenbu* (all)	*issei-ni* (all and simultaneously)	*saku* (be out)
	oya (parent)	*chichi* (father)	*nado* (etc.)
	hoso nagai (fine long)	*himojou no nagai* (corded long)	*sita de* (tongue)

Table 10. Examples of pairs having a similar meaning

Preceding contexts	Differing parts		Succeeding contexts
23 *do* (23 degrees)	27 (27)	26 (26)	*hun no isen* (minutes latitude)
	hori (a moat)	*ike* (a pond)	*ya* (or)
kaijou no (at sea)	*kokubou* (national defense)	*bouei, kougeki* (defense, attack)	*wo*
	ookina (big)	*hijouna* (extraordinary)	*rieki* (profit)

(= 1 − 0.342). When we considered the pairs that were hypernyms or similar expressions to be correct also, the precision decreased if we excluded cases where a difference on one side was a null expression. We explain this as follows. Consider Table 9, which shows expressions having a hypernym relationship. When a difference on one side is a null expression and the difference on the other side is an expression having an effect of expanding the region of a meaning, such as *nado* "etc." or *ijou* "more than", the difference pair relationship is that the difference which is a null expression is a hyponym of the other difference. When a difference on one side is a null expression and the difference on the other side is an expression restricting or decreasing the region of a meaning, the difference pair relationship is that the difference which is a null expression is a hypernym of the other difference. Therefore, when a difference on one side is a null expression, the pair of differences is likely to have a hypernym relation and the precision when including cases where a difference on one side is a null expression will increase.

When we considered pairs that were hypernyms or similar expressions to be correct also, the top X precision became extremely high. For example, the precision for the Top 500 of all data was 0.974 (= 1 − 0.026).

We can summarize our experimental results as follows.

– Our method using Equation 5 provided high precision for the Top X data. We also found that this method provided higher precision than the frequency method which is normally used. We can therefore effectively extract synonyms with high precision using our method.
– The number of extracted synonyms with our method was larger than that with Katoh's method. When we would like to extract more synonyms, we should not delete candidate synonym pairs through Katoh's Equation 6.
– In terms of synonym extraction, the precision was higher when we excluded cases where a difference on one side was a null expression than when we include these cases. However, in terms of also extracting hypernyms and similar expressions, the precision when we included cases where a difference on one side was a null expression was higher than when we excluded these cases.
– Our method, which extracts many synonyms, provides high precision for top X data. We obtained a precision rate of 0.764 for the top 500 data pairs and 0.220 for 500 randomly extracted data pairs when only synonyms were considered a correct answer. We obtained a precision rate of 0.974 for the top 500 data pairs and 0.722 for 500 randomly extracted data pairs when hypernyms and similar expressions were also considered correct answers.

5 Related Studies

Our approach was to extract synonyms by matching a pair of text sections sharing the same meaning. Other studies on automatic extraction using this approach include the following.

– Use of multiple sentences translated from the same original sentence
  Since sentences translated from the same original sentence should have the same meaning, synonyms can be extracted by matching those sentences. Barzilay and McKeown obtained synonyms by using this method [1]. Shimohata also obtained synonyms by using this method and then used the extracted synonyms to improve the performance of machine translation [10].
– Use of document pairs having the same content
  Documents from multiple newspaper publishing companies are gathered and pairs of documents having the same content are extracted. By matching these pairs of documents, synonyms are extracted. Shinyama et al. obtained synonyms using this method [11]. They extracted document pairs having the same content by using proper nouns appearing in the documents.
– Use of pairs consisting spoken data and corresponding written data

Murata et al. used presentations at academic conferences as spoken data and the corresponding papers as written data. They obtained synonyms and rewriting rules for paraphrasing between spoken language and written language by matching spoken data and the corresponding written data. They also performed paraphrasing between the spoken and written language [7].

– Use of parts having the same content in a document
  Murata et al. obtained rewriting rules and synonyms by matching a patent claim and its embodiment [9]. A patent claim and its embodiment share the same meaning, so such a pair can be used to obtain synonyms.
– Use of pairs consisting of an original sentence and a summarized version
  An original sentence and its summarized version share the same meaning, so we can extract synonyms by matching them. Katoh et al. used this method and obtained rewriting rules and synonyms for summarization [3].

As described above, many studies have been done on extracting synonyms by matching sentences or texts having the same meaning. Our method of using Equation 5 should be useful for such studies because it is convenient and effective.

6 Conclusion

Studies on both paraphrasing and paraphrase extraction are important in various research fields such as sentence generation, summarization, and question-answering. In our current work, we have studied methods of automatic paraphrase extraction based on matching definitions of the same word in two dictionaries.

Through our experiments, we confirmed that our proposed method using Equation 5 provided higher precision for the top pairs than other existing methods. This method should therefore be useful and effective, for application in other studies on automatic synonym extraction.

In our experiments, the estimated number of synonyms extracted with our method was 15,000, a much larger number than were extracted with Katoh's method. Although there have been studies on the extraction of synonym expressions from text pairs having the same content, no studies have been reported where definition sentences in multiple dictionaries were used as text pairs having the same content and synonym expressions were extracted from them. This paper has explained how many synonym expressions can be extracted from such definition sentences.

For synonym extraction, the precision was higher when we excluded cases where a difference on one side was a null expression. However, when we considered pairs that were hypernyms and similar expressions, as well as synonym pairs, to be correct, the precision was higher if we did not exclude cases where a difference on one side was a null expression.

With our method, we obtained precision rates of 0.764 for the top 500 data and 0.220 for 500 randomly extracted data when we considered only synonyms to be a correct answer. We obtained corresponding precision rates of 0.974 and 0.722, though, when we considered hypernyms and similar expressions to also be correct answers.

References

1. Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *39th Annual Meeting of the Association of the Computational Linguistics*, pages 50–57, 2001.
2. IWPT committe. Nlprs'2001 workshop on automatic paraphrasing: Theories and applications. 2001.
3. Naoto Katoh and Noriyoshi Uratani. A new approach to acquiring linguistic knowledge for locally summarizing Japanese news sentences. *Journal of Natural Language Processing*, 6(7), 1999. (in Japanese).
4. Sadao Kurohashi and Makoto Nagao. *Japanese Morphological Analysis System JUMAN version 3.6*. Department of Informatics, Kyoto University, 1998. (in Japanese).
5. Masaki Murata. NLP using DIFF — use of convenient tool for detecting differences, MDIFF —. *Journal of Natural Language Processing*, 9(2), 2002. (in Japanese).
6. Masaki Murata and Hitoshi Isahara. Universal model for paraphrasing — using transformation based on a defined criteria —. In *NLPRS'2001 Workshop on Automatic Paraphrasing: Theories and Applications*, 2001.
7. Masaki Murata and Hitoshi Isahara. Automatic extraction of differences between spoken and written languages, and automatic translation from the written to the spoken language. In *LERC 2002*, 2002.
8. Masaki Murata and Hitoshi Isahara. Using the diff command for natural language processing. 2002. http://arxiv.org/abs/cs.CL/ 0208020.
9. Masaki Murata and Hitoshi Isahara. Using the diff command in patent documents. *Proceedings of the Third NTCIR Workshop (PATENT)*, 2002.
10. Mitsuo Shimohata. Acquiring paraphrases from corpora and its application to machine translation. *Doctor's Thesis*, 2004. NAIST-IS-DD0261014.
11. Yusuke Shinyama and Satoshi Sekine. Paraphrase acquisition for information extraction. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, 2003.
12. Tetsuro Takahashi, Kozo Nawata, Kentaro Inui, and Yuji Matsumoto. Effect of structural matching and paraphrasing in question answering. *IEICE Transactions on Information and Systems*, E86–D(9):1677–1685, 2003.