Paraphrasing Predicates from Written Language to Spoken Language Using the Web

Nobuhiro Kaji and **Masashi Okamoto** and **Sadao Kurohashi**

Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

{kaji,okamoto,kuro}@kc.t.u-tokyo.ac.jp

Abstract

There are a lot of differences between expressions used in written language and spoken language. It is one of the reasons why speech synthesis applications are prone to produce unnatural speech. This paper represents a method of paraphrasing unsuitable expressions for spoken language into suitable ones. Those two expressions can be distinguished based on the occurrence probability in written and spoken language corpora which are automatically collected from the Web. Experimental results indicated the effectiveness of our method. The precision of the collected corpora was 94%, and the accuracy of learning paraphrases was 76 %.

1 Introduction

Information can be provided in various forms, and one of them is speech form. Speech form is familiar to humans, and can convey information effectively (Nadamoto et al., 2001; Hayashi et al., 1999). However, little electronic information is provided in speech form so far. On the other hand, there is a lot of information in text form, and it can be transformed into speech by a speech synthesis. Therefore, a lot of attention has been given to applications which uses speech synthesis, for example (Fukuhara et al., 2001).

In order to enhance such applications, two problems need to be resolved. The first is that current speech synthesis technology is still insufficient and many applications often produce speech with unnatural accents and intonations. The second one is that there are a lot of differences between expressions used in written language and spoken language. For example, Ohishi indicated that difficult words and compound nouns are more often used in written language than in spoken language (Ohishi, 1970). Therefore, the applications are prone to produce unnatural speech, if their input is in written language.

Although the first problem is well-known, little attention has been given to the second one. The reason why the second problem arises is that the input text contains Unsuitable Expressions for Spoken language (UES). Therefore, the problem can be resolved by paraphrasing UES into Suitable Expression for Spoken language (SES). This is a new application of paraphrasing. There are no similar attempts, although a variety of applications have been discussed so far, for example question-answering (Lin and Pantel, 2001; Hermjakob et al., 2002; Duclaye and Yvon, 2003) or text-simplification (Inui et al., 2003).

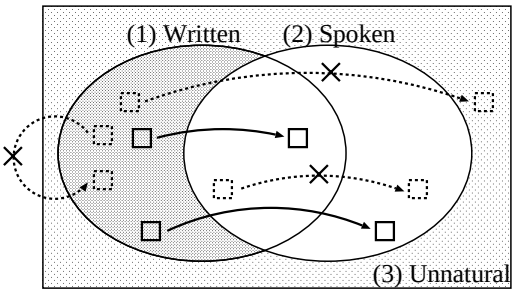

Figure 1: Paraphrasing UES into SES

Figure 1 illustrates paraphrasing UES into SES. In the figure, three types of expressions are shown: (1) expressions used in written language, (2) expressions used in spoken language, and (3) unnatural expressions. The overlap between two circles represents expressions used both in written language and spoken language. UES is the shaded portion: unnatural expressions, and expressions used only in written language. SES is the non-shaded portion. The arrows represent paraphrasing UES into SES, and other paraphrasing is represented by broken arrows. Paraphrasing unnatural expressions is not considered, since such expressions are not included in the input text. The reason why unnatural expressions are taken into consideration is that paraphrasing into such expressions should be avoided.

In order to paraphrase UES into SES, this paper proposes a method of learning paraphrase pairs in the form of 'UES → SES'. The key notion of the method is to distinguish UES and SES based on the occurrence probability in written and spoken language corpora which are automatically collected from the Web. The procedure of the method is as follows:[1]

(step 1) Paraphrase pairs of predicates[2] are learned from a dictionary using a method proposed by (Kaji et al., 2002).

(step 2) Written and spoken language corpora are automatically collected from the Web.

(step 3) From the paraphrase pairs learned in step 1, those in the form of 'UES → SES' are selected using the corpora.

This paper deals with only paraphrase pairs of predicates, although UES includes not only predicates but also other categories such as nouns.

This paper is organized as follows. In Section 2 related works are illustrated. Section 3 summarizes the method of Kaji et al. In Section 4, we describe the method of collecting corpora form the Web and report the experimental result. In Section 5, we describe the method of selecting suitable paraphrases pairs and the experimental result. Our future work is described in Section 6, and we conclude in Section 7.

2 Related Work

Paraphrases are different expressions which convey the same or almost the same meaning. However, there are few paraphrases that have exactly the same meaning, and almost all have subtle differences such as style or formality etc. Such a difference is called a connotational difference. This paper addresses one of the connotational differences, that is, the difference of whether an expression is suitable or unsuitable for spoken language.

Although a large number of studies have been made on learning paraphrases, for example (Barzilay and Lee, 2003), there are only a few studies which address the connotational difference of paraphrases. One of the studies is a series of works by Edmonds et al. and Inkpen et al (Edmonds and Hirst, 2002; Inkpen and Hirst, 2001). Edmonds et al. proposed a computational model which represents the connotational difference, and Inkpen et al. showed that the parameters of the model can be learned from a synonym dictionary. However, it is doubtful whether the connotational difference between paraphrases is sufficiently described in such a lexical resource. On the other hand, Inui et al. discussed readability, which is one of the connotational differences, and proposed a method of learning readability ranking model of paraphrases from a tagged corpus (Inui and Yamamoto, 2001). The tagged corpus was built as follows: a large amount of paraphrase pairs were prepared and annotators tagged them according to their readability. However, they focused only on syntactic paraphrases. This paper deals with lexical paraphrases.

There are several works that try to learn paraphrase pairs from parallel or comparable corpora (Barzilay and McKeown, 2001; Shinyama et al., 2002; Barzilay and Lee, 2003; Pang et al., 2003). In our work, paraphrase pairs are not learned from corpora but learned from a dictionary. Our corpora are neither parallel nor comparable, and are used to distinguish UES and SES.

There are several studies that compare two corpora which have different styles, for example, written and spoken corpora or British and American English corpora, and try to find expressions unique to either of the styles (Kilgarriff, 2001). However, those studies did not deal with paraphrases.

Bulyko et al. also collected spoken language corpora from the Web (Bulyko et al., 2003). The method of Bulyko et al. used N-grams in a training corpus and is different from ours (the detail of our method is described in Section 4).

In respect of automatically collecting corpora which have a desired style, Tambouratzis et al. proposed a method of dividing Modern Greek corpus into Demokiti and Katharevoua, which are variations of Modern Greek (Tambouratzis et al., 2000).

3 Learning Predicate Paraphrase Pairs

Kaji et al. proposed a method of paraphrasing predicates using a dictionary (Kaji et al., 2002). For example, when a definition sentence of '*chiratsuku* (to shimmer)' is '*yowaku hikaru* (to shine faintly)', his method paraphrases (1a) into (1b).

(1)	a.	*ranpu-ga*	*chiratsuku*	
		a lamp	to shimmer	
	b.	*ranpu-ga*	*yowaku*	*hikaru*
		a lamp	faintly	to shine

As Kaji et al. discussed, this dictionary-based paraphrasing involves three difficulties: word sense ambiguity, extraction of the appropriate paraphrase from a definition sentence, transformation of postposition[3]. In order to solve those difficulties, he proposed a method based on case frame alignment.

If paraphrases can be extracted from the definition sentences appropriately, paraphrase pairs can be learned. We extracted paraphrases from definition sentences using the

[1] Note that this paper deals with Japanese.

[2] A predicate is a verb or an adjective.

[3] Japanese noun is attached with a postposition.

method of Kaji et al. However, it is beyond the scope of this paper to describe his method as a whole. Instead, we represent an overview and show examples.

		(predicate)	(definition sentence)	
(2)	a.	*chiratsuku*	[*kasukani*	*hikaru*]
		to shimmer	faintly	to shine
			to shine faintly	
	b.	*chokinsuru*	[*okane-wo*	*tameru*]
		to save money	money	to save
			to save money	
	c.	*kansensuru*	*byouki-ga*	[*utsuru*]
		to be infected	disease	to be infected
			to be infected with a disease	

In almost all cases, a headword of a definition sentence of a predicate is also a predicate, and the definition sentence sometimes has adverbs and nouns which modify the head word. In the examples, headwords are '*hikaru* (to shine)', '*tameru* (to save)', and '*utsuru* (to be infected)'. The adverbs are underlined, the nouns are underlined doubly, paraphrases of the predicates are in brackets. The headword and the adverbs can be considered to be always included in the paraphrase. On the other hand, the nouns are not, for example 'money' in (2b) is included but 'disease' in (2c) is not. It is decided by the method of Kaji et al. whether they are included or not.

The paraphrase includes one noun at most, and is in the form of 'adverb$*$ noun+ predicate'[4]. Hereafter, it is assumed that a paraphrase pair which is learned is in the form of 'predicate $\rightarrow$ adverb$*$ noun+ predicate'. The predicate is called *source*, the 'adverb$*$ noun+ predicate' is called *target*.

We used *reikai-shougaku-dictionary* (Tadika, 1997), and 5,836 paraphrase pairs were learned. The main problem dealt with in this paper is to select paraphrase pairs in the form of 'UES $\rightarrow$ SES' from those 5,836 ones.

4 Collecting Written and Spoken Language Corpora from the Web

We distinguish UES and SES (see Figure 1) using the occurrence probability in written and spoken language corpora. Therefore, large written and spoken corpora are necessary. We cannot use existing Japanese spoken language corpora, such as (Maekawa et al., 2000; Takezawa et al., 2002), because they are small.

Our solution is to automatically collect written and spoken language corpora from the Web. The Web contains various texts in different styles. Such texts as news articles can be regarded as written language corpora, and such texts as chat logs can be regarded as spoken language corpora. Since we do not need information such as accents or intonations, speech data of real conversations is not always required.

This papepr proposes a method of collecting written and spoken language corpora from the Web using interpersonal expressions (Figure 2). Our method is as follows. First, a corpus is created by removing useless parts such as html tags from the Web. It is called *Web corpus*. Note that the Web corpus consist of Web pages (hereafter page). Secondly, the pages are classified into three types (written language corpus, spoken language corpus, and ambiguous corpus) based on interpersonal expressions. And then, only written and spoken language copora are used, and the ambiguous corpus is abandoned. This is because:

- Texts in the same page tend to be described in the same style.
- The boundary between written and spoken language is not clear even for humans, and it is almost impossible to precisely classify all pages into written language or spoken language.

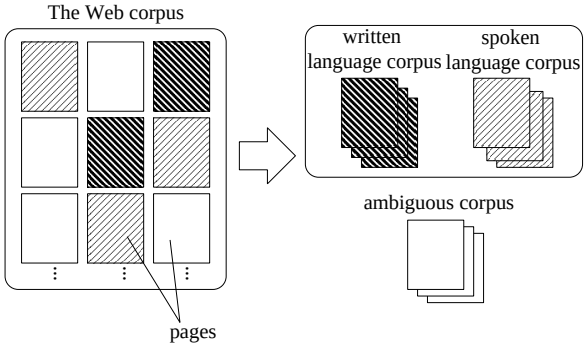

Figure 2: Collecting written and spoken language corpora

4.1 Interpersonal expressions

Each page in the Web corpus is classified based on interpersonal expressions.

Spoken language is often used as a medium of information which is directed to a specific listener. For example, face-to-face communication is one of the typical situations in which spoken language is used. Due to this fact, spoken language tends to contain expressions which imply an certain attitude of a speaker toward listeners, such as familiarity, politeness, honor or contempt etc. Such an expression is called *interpersonal expression*. On the other hand, written language is mostly directed to unspecific readers. For example, written language is often used in news articles or books or papers etc. Therefore, interpersonal expressions are not used so frequently in written language as in spoken language.

[4] $*$ means zero or more, and + means one or more.

Among interpersonal expressions, we utilized familiarity and politeness expressions. The familiarity expression is one kind of interpersonal expressions, which implies the speaker's familiarity toward the listener. It is represented by a postpositional particle such as '*ne*' or '*yo*' etc. The following is an example:

(3) *watashi-wa* *ureshikatta* *yo*
I was happy (**familiarity**)

I was happy

(3) implies familiarity using the postpositional particle '*yo*'.

The politeness expression is also one kind of interpersonal expressions, which implies politeness to the listener. It is represented by a postpositional particle. For example:

(4) *watashi-wa* *eiga-wo* *mi* *masu*
I a movie to watch (**politeness**)

I watch a movie

(4) implies politeness using the postpositional particle '*masu*'.

Those two interpersonal expressions often appear in spoken language, and are easily recognized as such by a morphological analyzer and simple rules. Therefore, a page in the Web corpus can be classified into the three types based the following two ratios.

- Familiarity ratio (F-ratio):

$$\frac{\text{\# of sentences which include familiarity expressions}}{\text{\# of all the sentences in the page}}$$

- Politeness ratio (P-ratio):

$$\frac{\text{\# of sentences which include politeness expressions}}{\text{\# of all the sentences in the page.}}$$

4.2 Algorithm

After the Web corpus is processed by a Japanese morphological analyzer (JUMAN)[5], sentences which include familiarity or politeness expressions are recognized in the following manner in order to calculate F-ratio and P-ratio. If a sentence has one of the following six postpositional particles, it is considered to include the familiarly expression.

ne, *yo*, *wa*, *sa*, *ze*, *na*

A sentence is considered to include the politeness expression, if it has one of the following four postpositional particles.

desu, *masu*, *kudasai*, *gozaimasu*

[5]http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman-e.html

If F-ratio and P-ratio of a page are very low, the page is in written language, and vice versa. We observed a part of the Web corpus, and empirically decided the rules illustrated in Table 1. If F-ratio and P-ratio are equal to 0, the page is classified as written language. If F-ratio is more than 0.2, or if F-ratio is more than 0.1 and P-ratio is more than 0.2, the page is classified as spoken language. The other pages are regarded as ambiguous.

Table 1: Page classification rules

F-ratio $= 0$ P-ratio $= 0$	$\rightarrow$ Written language
F-ratio > 0.2 or F-ratio > 0.1 P-ratio > 0.2	$\rightarrow$ Spoken language
Otherwise	$\rightarrow$ Ambiguous

4.3 Evaluation

The Web corpus we prepared consists of 660,062 pages and contains 733M words. Table 2 shows the size of the written and spoken language corpora which were collected from the Web corpus.

Table 2: The size of the corpora

	# of pages	# of words
The Web corpus	660,062	733M
Written language corpus	80,685	77M
Spoken language corpus	73,977	113M

Size comparison The reason why written and spoken language corpora were collected from the Web is that Japanese spoken language corpora available are too small. As far as we know, the biggest Japanese one is Spontaneous Speech Corpus of Japanese, which contains 7M words (Maekawa et al., 2000). Our corpus is about ten times as big as Spontaneous Speech Corpus of Japanese.

Precision of our method What is important for our method is not recall but precision. Even if the recall is not high we can collect large corpora, because the Web corpus is very huge. However, if the precision is low, it is impossible to collect corpora with high quality.

240 pages of the written and spoken language corpora were extracted at random, and the precision of our method was evaluated. The 240 pages consist of 125 pages collected as written language corpus and 115 pages collected as spoken language corpus. Two judges (hereafter judge 1 and 2) respectively assessed how many of the 240 pages were classified properly.

The result is shown in Table 3. The judge 1 identified 228 pages as properly classified ones; the judge 2 identified 221 pages as properly classified ones. The average precision of the total was 94% (=228+221/240+240) and we can say that our corpora have sufficient quality.

Table 3: # of pages properly collected

	Judge 1	Judge 2
Written language corpus	119/125	110/125
Spoken language corpus	109/115	111/115
Total	**228/240**	**221/240**

Discussion Pages which were inappropriately collected were examined, and it was found that lexical information is useful in order to properly classify them. (5) is an example which means 'A new software is exciting'.

(5) *atarashii* *sohuto-ha* *wakuwakusuru*
new software exiting

(5) is in spoken language, although it does not include any familiarity and politeness expressions. This is because of the word '*wakuwakusuru*', which is informal and means 'exiting'.

On way to deal with such pages is to use words characteristic of written or spoken language. Such words will be able to be gathered form our written and spoken language corpora. It is our future work to improve the quality of our corpora in an iterative way.

5 Paraphrase Pair Selection

A paraphrase pair we want is one in which the source is UES and the target is SES. From the paraphrase pairs learned in Section 3, such paraphrase pairs are selected using the written and spoken language corpora.

Occurrence probabilities (OPs) of expressions in the written and spoken language corpora can be used to distinguish UES and SES. This is because:

- An expression is likely to be UES if its OP in spoken language corpora is very low.
- An expression is likely to be UES, if its OP in written language corpora is much higher than that in spoken language corpora.

For example, Table 4 shows OP of '*jikaisuru*'. It is a difficult verb which means 'to admonish oneself', and rarely used in a conversation. The verb '*jikaisuru*' appeared 14 times in the written language corpus, which contains 6.1M predicates, and 7 times in the spoken language corpus, which contains 11.7M predicates. The OP of *jikaisuru* in spoken language corpus is low, compared with that in written language corpus. Therefore, we can say that '*jikaisuru*' is UES.

Table 4: Occurrence probability of '*jikaisuru*'

	written language corpus	spoken language corpus
# of *jikaisuru*	14	7
# of predicates	6.1M	11.7M
OP of *jikaisuru*	14/6.1M	7/11.7M

The paraphrase pair we want can be selected based on the following four OPs.

(1) OP of source in the written language corpus

(2) OP of source in the spoken language corpus

(3) OP of target in the written language corpus

(4) OP of target in the spoken language corpus

The selection can be considered as a binary classification task: paraphrase pairs in which source is UES and target is SES are treated as positive, and others are negative. We propose a method based on Support Vector Machine (Vapnik, 1995). The four OPs above are used as features.

5.1 Feature calculation

The method of calculating OP of an expression e ($= OP(e)$) in a corpus is described. According to the method, those four features can be calculated. The method is broken down into two steps: counting the frequency of e, and calculation of $OP(e)$ using the frequency.

Frequency After a corpus is processed by the Japanese morphological analyzer (JUMAN) and the parser (KNP)[6], the frequency of *e* ($F(e)$) is counted. Although the frequency is often obvious from the analysis result, there are several issues to be discussed.

The frequency of a predicate is sometimes quite different from that of the same predicate in the different voice. Therefore, the same predicates which have different voice should be treated as different predicates.

As already mentioned in Section 3, the form of source is 'predicate' and that of target is 'adjective$*$ noun+ predicate'. If *e* is target and contains adverbs and nouns, it is difficult to count the frequency because of the sparse data problem. In order to avoid the problem, an approximation that the adverbs are ignored is used. For example, the frequency of 'run fast' is approximated by that of 'run'. We did not ignore the noun because of the following reason. As a noun and a predicate forms an idiomatic phrase more often than an adverb and a predicate, the meaning of such idiomatic phrase completely changes without the noun.

[6] http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp-e.html

If the form of target is 'adverb$*$ noun predicate', the frequency is approximated by that of 'noun predicate', which is counted based on the parse result. However, generally speaking, the accuracy of Japanese parser is low compared with that of Japanese morphological analyzer; the former is about 90% while the latter about 99%. Therefore, only reliable part of the parse result is used in the same way as Kawahara et al. did. See (Kawahara and Kurohashi, 2001) for the details. Kawahara et al. reported that 97% accuracy is achieved in the reliable part.

Occurrence probability In general, $OP(e)$ is defined as:

$$OP(e) = F(e)/ \text{ \# of expressions in a corpus.}$$

$F(e)$ tends to be small when e contains a noun, because only a reliable part of the parsed corpus is used to count $F(e)$. Therefore, the value of the denominator '# of expressions in a corpus' should be changed depending on whether e contains a noun or not. The occurrence probability is defined as follows:

if e does not contain any nouns

$$OP(e) = F(e)/ \text{ \# of predicates in a corpus.}$$

otherwise

$$OP(e) = F(e)/ \text{ \# of noun-predicates in a corpus.}$$

Table 5 illustrates # of predicates and # of noun-predicates in our corpora.

Table 5: # of predicates, and # of noun-predicates

	# of predicates	# of noun-predicates
written language corpus	6.1M	1.5M
spoken language corpus	11.7M	1.9M

5.2 Evaluation

The two judges built a data set, and 20-hold cross validation was used.

Data set 267 paraphrase pairs were extracted at random form the 5,836 paraphrase pairs learned in section 3. Two judges independently tagged each of the 267 paraphrase pairs as positive or negative. Then, only such paraphrase pairs that were agreed upon by both of them were used as data set. The data set consists of 200 paraphrase pairs (70 positive pairs and 130 negative pairs).

Experimental result We implemented the system using Tiny SVM package[7].The Kernel function explored was the polynomial function of degree 2.

Using 20-hold cross validation, two types of feature sets (F-set1 and F-set2) were evaluated. F-set1 is a feature set of all the four features, and F-set2 is that of only two features: OP of source in the spoken language corpus, and OP of target in the spoken language corpus.

The results were evaluated through three measures: accuracy of the classification (positive or negative), precision of positive paraphrase pairs, and recall of positive paraphrase pairs. Table 6 shows the result. The accuracy, precision and recall of F-set1 were 76 %, 70 % and 73 % respectively. Those of F-set2 were 75 %, 67 %, and 69 %.

Table 6: Accuracy, precision and recall

	F-set1	F-set2
Accuracy	**76**%	75%
Precision	**70**%	67%
Recall	**73**%	69%

Table 7 shows examples of classification. The paraphrase pair (1) is positive example and the paraphrase pair (2) is negative, and both of them were successfully classified. The source of (1) appears only 10 times in the spoken language corpus, on the other hand, the source of (2) does 67 times.

Discussion It is challenging to detect the connotational difference between lexical paraphrases, and all the features were not explicitly given but estimated using the corpora which were prepared in the unsupervised manner. Therefore, we think that the accuracy of 76 % is very high.

The result of F-set1 exceeds that of F-set2. This indicates that comparing $OP(e)$ in the written and spoken language corpus is effective.

Calculated $OP(e)$ was occasionally quite far from our intuition. One example is that of '*kangekisuru*', which is a very difficult verb that means 'to watch a drama'. Although the verb is rarely used in real spoken language, its occurrence probability in the spoken language corpus was very high: the verb appeared 9 times in the written language corpus and 69 times in the spoken language corpus. We examined those corpora, and found that the spoken language corpus happens to contain a lot of texts about dramas. Such problems caused by biased topics will be resolved by collecting corpora form larger Web corpus.

[7] http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

Table 7: Successfully classified paraphrase pairs

Paraphrase pair			Occurrence probabilities			
			source		target	
			written language	spoken language	written language	spoken language
(1)	*denraisuru* to descend	→	*tsutawaru* to be transmitted	43/6.1M	10/11.7M	1,927/6.1M	4,213/11.7M
(2)	*hebaru* to be tired out	→	*hetohetoni tsukareru* to be exhausted	18/6.1M	67/11.7M	1,026/6.1M	7,829/11.7M

6 Future Work

In order to estimate more reliable features, we are going to increase the size of our corpora by preparing larger Web corpus.

Although the paper has discussed paraphrasing from the point of view that an expression is UES or SES, there are a variety of SESs such as slang or male/female speech etc. One of our future work is to examine what kind of spoken language is suitable for such a kind of application that was illustrated in the introduction.

This paper has focused only on paraphrasing predicates. However, there are other kinds of paraphrasing which are necessary in order to paraphrase written language text into spoken language. For example, paraphrasing compound nouns or complex syntactic structure is the task to be tackled.

7 Conclusion

This paper represented the method of learning paraphrase pairs in which source is UES and target is SES. The key notion of the method is to identify UES and SES based on the occurrence probability in the written and spoken language corpora which are automatically collected from the Web. The experimental result indicated that reliable corpora can be collected sufficiently, and the occurrence probability calculated from the corpora is useful to identify UES and SES.

References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57.

Ivan Bulyko, Mari Ostenforf, and Andreas Stolcke. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proceedings of HLT-NAACL2003*, pages 7–9.

Florence Duclaye and FranÇois Yvon. 2003. Learning paraphrases to improve a question-answering system. In *Proceedings of the 10th Conference of EACL Workshop Natural Language Processing for Question-Answering*.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Tomohiro Fukuhara, Toyoaki Nishida, and Shunsuke Uemura. 2001. Public opinion channel: A system for augmenting social intelligence of a community. In *Workshop notes of the JSAI-Synsophy International Conference on Social Intelligence Design*, pages 22–25.

Masaki Hayashi, Hirotada Ueda, Tsuneya Kurihara, Michiaki Yasumura, Mamoru Douke, and Kyoko Ariyasu. 1999. Tvml (tv program making language) - automatic tv program generation from text-based script -. In *ABU Technical Review*.

Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural language based reformulation resource and web exploitation for question answering. In *Proceedings of TREC 2002 Conference*.

Diana Zaiu Inkpen and Graeme Hirst. 2001. Building a lexical knowledge-base of near-synonym differences. In *Proceedings of Workshop on WordNet and Other Lexical Sources*, pages 47–52.

Kentaro Inui and Satomi Yamamoto. 2001. Corpus-based acquisition of sentence readability ranking models for deaf people. In *Proceedings of NLPRS 2001*.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: A project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16.

Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato. 2002. Verb paraphrase based on case frame alignment. In *Proceedings of ACL 2002*, pages 215–222.

Daisuke Kawahara and Sadao Kurohashi. 2001. Japanese case frame construction by coupling the verb and its closest case component. In *Proceedings of HLT 2001*, pages 204–210.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Journal of Natural Language Engneering*, 7(4):343–360.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *Proceedings of LREC 2000*, pages 947–952.

Akiyo Nadamoto, Hiroyuki Kondo, and Katsumi Tanaka. 2001. Webcarousel: Restructuring web search results for passive viewing in mobile environments. In *7th International Conference on Database Systems for Advanced Applications*, pages 164–165.

Hatsutaroh Ohishi, editor. 1970. *Hanashi Kotoba(Spoken Language)*. Bunkacho.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating sentences. In *Proceedings of HLT-NAACL 2003*.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT 2002*.

Jyunichi Tadika, editor. 1997. *Reikai Shougaku Kokugojiten (Japanese dictionary for children)*. Sanseido.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, pages 147–152.

George Tambouratzis, Stella Markantonatou, Nikolaos Hairetakis, Marina Vassiliou, Dimitrios Tambouratzis, and George Carayannis. 2000. Discriminating the registers and styles in the modern greek language. In *Proceedings of Workshop on Comparing Corpora 2000*.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.