Paraphrasing of Chinese Utterances

Yujie Zhang[*]
Communications Research Laboratory
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289 Japan
yujie@crl.go.jp

Kazuhide Yamamoto
ATR Spoken Language Translation Research Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
yamamoto@fw.ipsj.or.jp

Abstract

One of the key issues in spoken language translation is how to deal with unrestricted expressions in spontaneous utterances. This research is centered on the development of a Chinese paraphraser that automatically paraphrases utterances prior to transfer in Chinese-Japanese spoken language translation. In this paper, a pattern-based approach to paraphrasing is proposed for which only morphological analysis is required. In addition, a pattern construction method is described through which paraphrasing patterns can be efficiently learned from a paraphrase corpus and human experience. Using the implemented paraphraser and the obtained patterns, a paraphrasing experiment was conducted and the results were evaluated.

1 Introduction

In spoken language translation one of the key issues is how to deal with unrestricted expressions in spontaneous utterances. To resolve this problem, we have proposed a paraphrasing approach in which the utterances are automatically paraphrased prior to transfer (Yamamoto et al., 2001; Yamamoto, 2002). The paraphrasing process aims to bridge the gap between the unrestricted expressions in the input and the limited expressions that the transfer can translate. In fact, paraphrasing actions are often seen in daily communication. When a listener cannot understand what a speaker said, the speaker usually says it again using other words, i.e., he paraphrases. In a Chinese-Japanese spoken language translation system, the pre-processing of Chinese utterances is involved and we attempt to apply a paraphrasing approach. This paper is focused on the paraphrasing of Chinese utterances.

Some cases of paraphrasing research with certain targets have been reported. For example, there has been work on rewriting the source language in machine translation with a focus on reducing syntactic ambiguities (Shirai et al., 1993), research on paraphrasing paper titles with a focus on transforming syntactic structures to achieve readability (Sato, 1999), and research on paraphrasing Japanese in summarization with a focus on transforming a noun modifier into a noun phrase (Kataoka et al., 1999). We have reported some research on Chinese paraphrasing (Zhang and Yamamoto, 2001; Zhang et al., 2001; Zong et al., 2001). The techniques of paraphrasing natural language can be applied not only to the pre-processing of machine translation but also to information retrieval and summarization.

2 Goals and Approach

In the pre-processing stage of translation, Chinese paraphrasing focuses on

(1) transforming the expressions of spoken language into formal expressions,

(2) reducing syntactic and semantic ambiguities,

(3) generating as many different expressions as possible in order to include expressions that can be translated by the transfer, and

(4) paraphrasing the main constituents of the utterance in case the paraphrasing of the whole utterance has no effect.

The aim of paraphrasing types (1), (2) and (4) is to simplify the expressions of utterances, and that of paraphrasing type (3) is to increase the

[*] This work was done when the author stayed at ATR Spoken Language Translation Research Laboratories.

variations of utterances. At present, we focus on paraphrasing types (1), (2) and (3).

Paraphrasing is a process that automatically generates new expressions that have the same meaning as the input sentence. At first glance one would think that the problem could be resolved by separating it into two processes: the parsing process that analyzes the input sentence and obtains its meaning, and the generation process that generates sentences from the obtained meaning. However, this solution is not practicable for the following reasons.

- At present, the techniques of parsing and semantics analysis of the Chinese language are far below the level needed for application. When studying spoken language, research on parsing and research on semantics analysis are major themes themselves. For automatic paraphrasing, we should first determine what kind of analysis is required and then start to develop a parser or a semantics analyzer.
- Even if meanings can be obtained, goal (3) cannot be achieved if only one sentence is generated. Here, the demand that paraphrasing should generate multiple expressions is the most important. This focus is different from that of conventional sentence generation.

In fact, the paraphrasing can be conducted at many different levels, for instance, words, phrases, or larger constituents. Although the paraphrasing of such constituents is probably related to context, it is not true that paraphrasing is impossible without being able to understand the whole sentence (Kataoka et al., 1999).

The paraphrasing process encounters the following problems. (i) How to identify objects, i.e., which components of an input sentence will be paraphrased, (ii) how to generate new sentences, and (iii) how to ensure that the generated sentences have the same meaning as the input sentence. In order to avoid the large cost of syntax and semantics analysis, we propose a pattern-based approach to paraphrasing in which only morphological analysis is required. The focus is placed on how to generate as many different expressions as possible and how to get paraphrasing patterns from a paraphrasing corpus.

Table 1. Part of the part-of-speech tag set of the Penn Chinese Treebank

Symbol	Explanation
NN	common noun
NR	proper noun
PN	pronoun
DT	determiner
DEC	的 in a relative-clause
DEG	associative 的
M	measure word
JJ	other noun-modifier
VA	predicative adjective
VC	是
VE	有 as the main verb
VV	other verb
AD	adverb
P	preposition excl. 被 and 把
LC	localizer
CD	cardinal number
OD	ordinal number
SP	sentence-final particle
BA	把 in ba-construction
CC	coordinating conjunction

3 Paraphrasing Pattern

The paraphrase corpus of the spoken Chinese language consists of 20,000 original sentences and 44,480 paraphrases, one original sentence having at least two paraphrases (Zhang et al., 2001). The paraphrases were obtained by the manual rephrasing of the original sentences: words may be reordered, some words may be substituted with synonyms, or the syntactic structures may be changed. Such a paraphrase corpus contains the knowledge of how to generate paraphrases for one sentence. We intend to get paraphrasing patterns from the corpus. By pairing each paraphrase with its corresponding original sentence, 44,480 pairs were obtained. Hereafter, we call such pairs paraphrase pairs. Word segmentation and part-of-speech tagging were carried out on the paraphrase pairs. The part-of-speech tagger accepted the Penn Chinese Treebank tag set, which comprises 33 parts-of-speech (Xia, 2000). A part of the Penn Chinese Treebank tag set is shown in Table 1.

3.1 Extraction of Instances

For one paraphrase pair, the paraphrase may differ from its original sentence in one of the following paraphrasing phenomena: (1) word order, (2) substitution of synonyms, and (3) change of syntactic structure. For most paraphrase pairs, the paraphrases contain a mixture of the above phenomena. We need to classify the paraphrasing phenomena and learn the relative paraphrasing patterns. In this way, we can restrict the paraphrasing process to some language phenomena and summarize the changes in the information of the resultant paraphrases. The following paraphrasing phenomena were considered and related paraphrase pairs were extracted.

3.1.1 Word Order

Word order in the spoken Chinese is comparatively free. In the paraphrase corpus, quite a large proportion of the paraphrases is created by word reordering. We extracted the paraphrase pairs in which the morpheme number of the original sentence is equal to that of the paraphrase and each morpheme of the original sentence appears in the paraphrase and vice versa. One example is shown in 3-1.

[**3 – 1**] An extracted paraphrase pair.
Original: 请 /VV 再 /AD 打电话 /VV 给 /P 我 /PN 好 /VA 吗 /SP
(Please call me again, could you?)
Paraphrase: 请 /VV 再 /AD 给 /P 我 /PN 打电话 /VV 好 /VA 吗 /SP

Guided by the extracted paraphrase pair, we can in fact paraphrase the original sentence by reordering its words according to the word order of the paraphrase. The extracted paraphrase pairs of this kind provided instances for learning word order paraphrasing patterns.

3.1.2 Negative Expressions

In some paraphrase pairs, we observed that paraphrasing phenomena were related to negative expressions. For example, original sentences include negative words "不 (do not)" or "没 (did not)" , but their corresponding paraphrases appear as affirmative forms without these negative words. This fact implied that the sentences could be simplified by deleting the negative expressions. For this purpose, the paraphrase pairs were extracted in which the original sentences included the words "不" or "没" and the corresponding paraphrases did not. One example is shown in 3-2.

[**3 – 2**]
Original: 知 /VV 不 /AD 知道 /VV 我 /PN 的 /DEG 電話 /NN
(Do you know my telephone number?)
Paraphrase: 知道 /VV 我 /PN 的 /DEG 電話 /NN 吗 /SP

3.1.3 Expression of "把"

The Chinese language has a few grammatical markers. The particle "把" is one of such markers. The sentences with the form "S(subject) V(verb) O(object) C(complement)" may be changed into the form "S 把 O V C" by inserting the particle "把" (Zhang and Sato, 1999). The usage of "把" emphasizes the object by moving it before the verb. When the particle "把" is in a sentence, it is easier to identify the object. So the insertion of "把" will supply more information about syntactic structure and reduce syntactic ambiguities. Moreover, paraphrasing the sentences with particle "把" may be more exact because the identification of the object is more accurate. We extracted the paraphrase pairs in which the original sentences included the particle "把" and the corresponding paraphrases did not. See example 3-3 below.

[**3 – 3**]
Original: 这 /DT 张 /M 单子 /NN 请 /VV 您 /PN 填好 /VV
(Could you fill out this form, please.)
Paraphrase: 请 /VV 您 /PN 把 /BA 这 /DT 张 /M 单子 /NN 填好 /VV
(Could you make this form filled out, please.)

3.2 Automatic Generalization of Instances

Then we attempted to generalize the extracted instances in order to obtain paraphrasing patterns. For each extracted paraphrase pair, the original sentence is generalized to make the matching part of the pattern, and the paraphrase is generalized to make the generation part of the pattern. The matching part specifies the components that will be paraphrased as well as the context conditions. The generation part defines how to construct a paraphrase. When the constituted pattern is applied to one input sentence, if the input matches with the matching part, a new sentence will be generated according to the generation part.

In fact, the purpose of generalization is to get a regular expression from the original sentence and to get an operation expression containing substitutions from the paraphrase. As shown in 3-3, both the original sentence and the paraphrase are series of morphemes, and each morpheme consists of a part-of-speech and an orthographic expression. The important thing in paraphrasing is to maintain meaning. To what extent the series of morphemes will be generalized depends on each paraphrasing pair. First, parts-of-speech keep the syntactic information and therefore they should be kept. Second, orthographic expressions of verbs, auxiliary verbs, adverbs, etc., are important in deciding the main meaning of the sentence and therefore they should also be kept. The orthographic expressions of other categories, such as nouns, pronouns and numerals, can be generalized to an abstract level by replacing each orthographic expression with a wild card.

The pattern generalized from 3-3 is illustrated in 3-4. The left part is the matching part and the right part is the generation part. The lexical information may be an orthographic expression or a variable represented by symbol X_i. X_i in the matching part is in fact a wild card, which means it can match with any orthographic expression in the matching operation. X_i in the generation part defines a substitution operation.

[**3 − 4**] A generalized pattern.
这 /DT 张 /M X_1/NN 请 /VV X_2/PN 填好 /VV → 请 /VV X_2/PN 把 /BA 这 /DT 张 /M X_1/NN 填好 /VV

However, we found two problems in this kind of automatic generalization. The first is that restrictions on the patterns generalized from long sentences are too specific at the lexical level. In fact, the clauses and noun phrases used as modifiers have no effect on the considered paraphrasing phenomena and can be generalized further. The second is that some orthographic expressions with important meanings are generalized to wild cards, for instance, the numeral “多少 (how many)” may imply that the sentence is interrogative. Therefore, a method is needed to prevent some orthographic expressions from being automatically replaced with wild cards.

3.3 Semi-Automatic Generalization of Instances

Specifying which morphemes should be generalized and which orthographic expressions should be kept requires human experience. In order to integrate human experience into automatic generalization, we developed a semi-automatic generalization tool. The tool consists of description symbols and a transformation program. The description symbols are designed for people to define generalization information on instances, and the transformation program automatically transforms the defined instances into patterns. Three description symbols are defined as follows.

[ ]: This symbol is followed by a numeral and is used to enclose a sequence of morphemes. The enclosed part is a syntactic component, e.g., a noun phrase or a clause. Except for the part-of-speech of the last morpheme, the enclosed part will be replaced with a variable. In the Chinese language, the syntactic property of a sequence of words is most likely reflected in the last word, so we keep the part-of-speech of the last morpheme. The enclosed parts in the original sentence and the paraphrase denoted by the same numerals will be replaced with the same variables.

{ }: This symbol is used to enclose a morpheme. The orthographic expression of the morpheme will be kept. In this way, the lexical information of morphemes can be utilized to define the context. A few orthographic expressions can be defined inside one symbol so that words that can be paraphrased in the same way can be stored as one pattern.

⟨ ⟩: This symbol is used to enclose a morpheme. The orthographic expression of the morpheme will be replaced with a variable. In this way, the orthographic expressions of verbs or adverbs can also be generalized.

The usage of the symbols is explained in 3-5 and 3-6. Example 3-5 is a paraphrase pair in which description symbols are defined. Example 3-6 is the paraphrasing pattern generalized from 3-5.

[**3** − **5**] A defined instance.
Original: 请/VV 给/VV 我/PN 两/CD ⟨本/M⟩ [日语/NN 的/DEG]$_1$ [指南/NN 手册/NN]$_2$ (Could you give me two copies of the Japanese pamphlet, please?)
Paraphrase: [日语/NN 的/DEG]$_1$ [导游/NN 手册/NN]$_2$ 请/VV 给/VV 我/PN 两/CD ⟨本/M⟩

[**3** − **6**] The generalized pattern.
请/VV 给/VV X_1/PN X_2/CD X_3/M Y_1/DEG Y_2/NN → Y_1/DEG Y_2/NN 请/VV 给/VV X_1/PN X_2/CD X_3/M

X_i has the same meaning as that of 3-4. Y_1/DEG in the matching part implies that it can match with any sequence of morphemes in which the part-of-speech of the last morpheme is equal to DEG. Y_1/DEG in the generation part defines a substitution operation. Y_2/NN implies the same meaning, but the part-of-speech of the last morpheme is equal to NN. In addition to the automatic generalization for morphemes of category PN and CD, the defined "⟨本/M⟩" is also generalized to X_3/M. The defined "[指南(guide)/NN 手册/NN]$_2$" in Original and "[导游(tourist guide)/NN 手册/NN]$_2$" in Paraphrase are both generalized to Y_2/NN, although they are not exactly the same.

3.4 Construction of the Paraphrasing Patterns

Using the developed tool, we manually defined generalization information on the extracted paraphrase pairs and then obtained the following four groups of paraphrasing patterns through automatic transformation.

(1) 459 patterns of deleting negative expressions.

(2) 160 patterns of inserting "把".

(3) 160 patterns of deleting "把".

(4) 2,030 patterns of reordering words.

The patterns of (3) were obtained by reversing the matching part and the generation part of each pattern of (2).

4 Design of the Paraphrasing Process

In order to generate as many different expressions as possible, we designed a mechanism for applying different groups of paraphrasing patterns. As described in Section 2, the paraphrasing process can be roughly classified into simplification paraphrasing aimed at simplifying expressions, and diversity paraphrasing aimed at increasing variations. Bearing in mind that simplification paraphrasing can reduce syntactic and semantic ambiguities, we apply this type of paraphrasing first, and then apply diversity paraphrasing. Using this strategy, we anticipate that the accuracy of diversity paraphrasing will be higher because there will be fewer ambiguities in syntax and semantics. In the four groups of patterns obtained above, group (1) belongs to simplification paraphrasing, and the other groups belong to diversity paraphrasing.

For one input sentence, the procedure for applying the different groups of patterns is designed as follows.

```
(1) Make the input sentence the application data for all groups of patterns. Set group number $i = 1$.
(2) In the application of group $i$, get one pattern from the group and repeat step (2.1) to step (2.3).
    (2.1) Match the input with the matching part of the selected pattern. If the matching succeeds, generate a sentence according to the generation part of the pattern.
    (2.2) Make the generated sentence the application data for all groups $j$ ($i < j \leq 4$). (At present there are four groups of patterns.)
    (2.3) Get another pattern then go to step (2.1) until there are no patterns left in group $i$.
(3) Set $i = i + 1$ and go to step (2) until $i > 4$.
(4) When passing the generated sentences to the transfer, do not pass duplicated ones.
```

Using this procedure, the generated paraphrases can be passed to the transfer at any time of the paraphrasing process. If one of the paraphrases can be translated by the transfer, the paraphrasing process will be stopped. In addition, the generated paraphrases can be paraphrased further by the patterns of following groups, therefore more expressions are likely to be produced. Based on this design, a paraphraser was implemented.

5 Experiment and Evaluation

A paraphrasing experiment was carried out on the paraphrase corpus using the implemented paraphraser and the obtained patterns. In order to get the same effect as that of using open test data, each pattern was not applied to the sentence from which the pattern was generalized. For 45,110 test sentences, 4,908 test sentences (about 10.9%) were paraphrased. From the 4,908 test sentences, 8,183 paraphrases were generated and the average number of paraphrases for one test sentence was 1.66. The generated paraphrases were evaluated by Chinese natives from two viewpoints, i.e., naturalness and meaning-retaining, with their corresponding test sentences. As a result, 7,226 generated paraphrases were correct and an 88% accuracy was achieved. The experimental result is shown in Table 2.

Table 2. Result of Paraphrasing Experiment

# Test Sentences	45,110
# Paraphrased Test Sentences	4,908 (10.9%)
# Generated Paraphrases	8,183
# Correct Paraphrases	7,226 (88%)

Three examples of the paraphrasing results are given below.

[5 − 1]
Input: 能不能 帮我订到最早的班机呢?
(Could you reserve the earliest plane for me?)
Paraphrase 1: 能 帮我订到最早的班机吗?
Paraphrase 2: 可以 帮我订到最早的班机吗?

[5 − 2]
Input: 在这儿 可以 订 饭店 吗
(May I reserve a restaurant here?)
Paraphrase 1: 可以 在这儿订 饭店 吗
Paraphrase 2: 在这儿订 饭店可以 吗
Paraphrase 3: 饭店 在这儿 可以 订吗
Paraphrase 4: 饭店可以 在这儿订吗
Paraphrase 5: 饭店 在这儿订 可以 吗

[5 − 3]
Input: 风景漂亮的房间 请给我.
(A room with a nice view, please give me.)
Paraphrase 1: 请 把 风景漂亮的房间给我.
(Please arrange a room with a nice view for me.)
Paraphrase 2: 我 想要 风景漂亮的房间.
(I would like a room with a nice view.)
Paraphrase 3: 请给我 风景漂亮的房间.
(Please give me a room with a nice view.)
Paraphrase 4: 房间 请给我 风景漂亮的.
(As for room, please give me one with a nice view.)

In the Input of 5-1 there is an expression of repeated interrogation "能不能" that consists of an affirmation "能 (can)" and a negation "不能 (can not)". After the application of the patterns of deleting negative expressions, Paraphrase 1 and Paraphrase 2 were generated. Both paraphrases are in affirmative form and both are correct. From the Input of 5-2, five paraphrases were generated only by reordering words. Paraphrases 1, 2, 3, 4 and 5 are all correct. In the Input of 5-3, the order of the predicate "请给我" and the object "风景漂亮的房间" is inverted. After the patterns of inserting "把" were applied, Paraphrase 1 was obtained. Then, the patterns of deleting "把" were applied to this generated paraphrase and Paraphrase 2 and Paraphrase 3 were obtained. In Paraphrase 3 the common word order was recovered. Finally, the patterns of reordering words were applied to Paraphrases 1, 2, 3 and the input. Paraphrase 4 was obtained from Paraphrase 3. Paraphrases 1, 2, 3 and 4 are all correct.

From the experimental results we see that the proposed approach can realize the goal of simplifying the expressions of the inputs and increasing variations with a high level of accuracy. If one of the paraphrased results can be translated, we can say that the paraphraser is effective in the translation system.

Through the analysis of wrong results, we found two reasons for paraphrasing errors. One reason is that some constituents or modification relations are incorrectly recognized based on the obtained paraphrasing patterns. For example, when pattern 3-6 was applied to the sentence "请给我两个人住的房间 (A room for two people, please.)", the quantity phrase "两个 (two)" was wrongly recognized as modifying "人住的房间 (a room where people can live)”, whereas it in fact modifies the noun “人 (people)”. Because of this wrong recognition, the generated sentence was “人住的房间请给我两个 (Please give me two rooms where people can live.)”. Its meaning is different from the input and therefore the result is not correct. The other reason for paraphrasing errors is that there were errors in part-of-speech tagging. For example, the word “给” in “请给我包装一下这礼物 (Please gift-wrap this)” was tagged as a verb while it really acts as a preposition. The wrong tagging resulted in the wrong application of patterns.

6 Conclusion

In this paper, a pattern-based approach to the paraphrasing of Chinese utterances is proposed and a method of constructing paraphrasing patterns from a corpus is described. Based on the proposed approach and method, a paraphraser is implemented and four types of paraphrasing patterns are constructed. Also, a paraphrasing experiment is conducted and experimental results are reported. The proposed approach has the following advantages.

(1) Because only morphological analysis is required, it is easy to implement the paraphraser and the processing time is short.

(2) By using the developed semi-automatic generalization tool, paraphrasing patterns can be efficiently learned from a paraphrasing corpus and human experience. The patterns enhanced by human experience have a higher accuracy.

(3) The classification of paraphrasing phenomena in pattern learning makes it possible to restrict the paraphrasing process to some language phenomena. The mechanism of applying different types of patterns emphasizes how to raise the accuracy of paraphrasing and how to increase variations.

In this research, only four types of paraphrasing phenomena are involved. The coverage achieved using the current patterns is still low. In the next phase, we are going to use the proposed approach on other paraphrasing phenomena in order to be able to paraphrase more Chinese utterances.

Acknowledgement

This research was supported in part by the Telecommunications Advancement Organization of Japan. We would also like to thank Ms. Lan Yao for her help in our experiments.

References

Akira Kataoka, Shigeru Masuyama, and Kazuhide Yamamoto. 1999. Summarization by Shortening a Japanese Noun Modifier into Expression “A no B”. In *Proc. of NLPRS'99*, pages 409–414.

Satoshi Sato. 1999. Automatic Paraphrase of Technical Papers. *Transactions of Information Processing Society of Japan*, 40(7):2937–2945. (in Japanese).

Satoshi Shirai, Satoru Ikehara, and Tsukasa Kawaoka. 1993. Effects of Automatic Rewriting of Source Language within a Japanese to English MT System. In *Proc. of TMI '93*, pages 226–239.

Fei Xia. 2000. The Part-of-speech Tagging Guideline for the Penn Chinese Treebank (3.0). Available at `http://www.ldc.upenn.edu/ctb`.

Kazuhide Yamamoto, Satoshi Shirai, Masashi Sakamoto, and Yujie Zhang. 2001. Sandglass: Twin Paraphrasing Spoken Language Translation. In *Proc. of ICCPOL'01*, pages 154–159.

Kazuhide Yamamoto. 2002. Machine Translation by Interaction between Paraphraser and Transfer. In *Proc. of Coling 2002*.

Li Zhang and Haruhiko Sato. 1999. *Chinese Expression Grammar-28 Points*. Toho-Shoten. (in Japanese).

Yujie Zhang and Kazuhide Yamamoto. 2001. Analysis of Chinese Spoken Language for Automatic Paraphrasing. In *Proc. of ICCPOL'01*, pages 290–293.

Yujie Zhang, Chengqing Zong, Kazuhide Yamamoto, and Masashi Sakamoto. 2001. Paraphrasing Utterances by Reordering Words Using Semi-Automatically Acquired Patterns. In *Proc. of NLPRS'01*, pages 195–202.

Chengqing Zong, Yujie Zhang, Kazuhide Yamamoto, Masashi Sakamoto, and Satoshi Shirai. 2001. Approach to Spoken Chinese Paraphrasing Based on Feature Extraction. In *Proc. of NLPRS'01*, pages 551–556.