*Natural Language Engineering* **11** (4): 417–434. © 2005 Cambridge University Press
doi:10.1017/S1351324905003712 Printed in the United Kingdom


# *Paraphrasing spoken Chinese using a paraphrase corpus*

YUJIE ZHANG

*National Institute of Information and Communications Technology, 3-5, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan*
*e-mail*: yujie@nict.go.jp

KAZUHIDE YAMAMOTO

*Nagaoka University of Technology, Niigata 940-2188 Japan*
*e-mail*: yamamoto@fw.ipsj.or.jp

(*Received 7 March 2003; revised 1 October 2004*)

**Abstract**

One of the key issues in spoken-language translation is how to deal with unrestricted expressions in spontaneous utterances. We have developed a paraphraser for use as part of a translation system, and in this paper we describe the implementation of a Chinese paraphraser for a Chinese-Japanese spoken-language translation system. When an input sentence cannot be translated by the transfer engine, the paraphraser automatically transforms the sentence into alternative expressions until one of these alternatives can be translated by the transfer engine. Two primary issues must be dealt with in paraphrasing: how to determine new expressions, and how to retain the meaning of the input sentence. We use a pattern-based approach in which the meaning is retained to the greatest possible extent without deep parsing. The paraphrase patterns are acquired from a paraphrase corpus and human experience. The paraphrase instances are automatically extracted and then generalized into paraphrase patterns. A total of 1719 paraphrase patterns obtained using this method and an implemented paraphraser were used in a paraphrasing experiment. The results showed that the implemented paraphraser generated 1.7 paraphrases on average for each test sentence and achieved an accuracy of 88%.

## 1 Introduction

Since spoken language is the most convenient means of human communication, automatic spoken-language translation (SLT) has become a research topic of considerable importance. Spoken language is more difficult to process than written language, mainly because of the unrestricted expression in spontaneous utterances. Conventional machine translation systems designed for written language generally perform poorly when applied to spoken language. To increase the robustness of SLT systems, we previously proposed a translation paradigm in which a paraphraser is used to transform unrestricted expressions into formal expressions (Yamamoto, Shirai, Sakamoto and Zhang 2001; Yamamoto 2002). The problem of unrestricted expression within the source language can be overcome by reinforcing the monolingual

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

processing, as there are enough monolingual linguists and monolingual corpora available for the development of a monolingual paraphraser. This approach greatly reduces the processing load on the transfer engine. Of particular importance, it reduces the need for bilingual linguists and bilingual corpora during the development of the transfer engine. This paper describes the implementation of this approach in a Chinese paraphraser for Chinese-Japanese spoken-language translation.

Paraphrasing is an important aspect of daily communication. A speaker who is not understood often repeats the message using a different expression, i.e., by paraphrasing. Paraphrasing is defined as an alternative way a speaker can choose to "say the same thing" by using linguistic knowledge (Iordanskaja, Kittredge and Polguere 1991). In natural language processing (NLP), automatic paraphrasing is regarded as the process of transforming one expression into a different expression while retaining the meaning. Paraphrasing techniques have been introduced into many NLP tasks, such as machine translation (MT) and information retrieval and summarization. In MT, rewriting input sentences into a set of restricted expressions is a kind of paraphrasing (Shirai, Ikehara and Kawaoka 1993). The purpose is to reduce ambiguity in the syntax and semantics of the input sentences before translation. In information retrieval, queries are paraphrased into multiple expressions to enable searching of more documents that are relevant (Zukerman and Raskutti 2002). In summarization, sentences are paraphrased into different expressions to identify redundant expressions, which are then ignored (Kataoka, Masuyama and Yamamoto 1999).

A paraphrasing process can be applied at different levels; for example, at the lexicon level, a synonym is substituted for a word (Stede 1996; Zukerman and Raskutti 2002), at the syntax level, the syntactic structure of a phrase or a larger constituent are changed (Barzilay and McKeown 2002), and at the pragmatic level, the whole sentence is changed into another one where the two sentences have different denotations but the same connotations (Sato 1999).

Two issues arise in automatic paraphrasing. One is how to retain the meaning of the input sentence. Some approaches use parsing and semantic analysis to obtain the internal expression of the meaning of the input sentence and then generate new sentences from the internal expression (Dras 1997; Ramsay and Seville 2000). This kind of paraphrasing involves deep analysis and therefore requires analytical techniques developed well enough for practical use. Other approaches focus on the constituents to be paraphrased, such as a word or a phrase, and use the surface information (Kataoka, Masuyama and Yamamoto 1999). The second issue is how to acquire paraphrase knowledge. Significant research concerning this issue has been reported recently. For instance, Barzilay applied multi-sequence alignment to a parallel corpus (Barzilay and McKeown 2001), and Pang applied a syntax-based algorithm to semantically equivalent translation sets (Pang, Knight and Marcu 2003). The data used in both of these studies were corpora consisting of multiple English translations of the same source text.

In this paper, we first define our goals for paraphrasing and describe our approach to paraphrasing. Then, we describe our methods for automatically extracting paraphrase instances from a corpus in section 3 and generalizing the instances into patterns in section 4. In section 5, we describe the implementation of our paraphraser.

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

We describe our experiment for evaluating the paraphraser and discuss the results in section 6. We conclude in section 7 with a brief summary and a look at future work.

## 2 Our approach

In this section, we first describe our SLT system and then set out our goals for paraphrasing and describe our approach to paraphrasing.

### *2.1 Outline of SLT system*

The paradigm of our SLT system is as follows. The system consists of five modules: a speech recognizer, a source-language paraphraser, a transfer engine, a target-language paraphraser, and a speech synthesizer. The source-language paraphraser changes input expressions into ones more likely to be translated correctly by the transfer engine. Unless stated otherwise, we use "paraphraser" to refer to the source-language paraphraser. When a sentence cannot be translated by the transfer engine, the paraphraser transforms the sentence into an alternative expression and sends it back to the transfer engine for another attempt. This process is repeated until an expression can be translated. The five modules are now being developed independently.

The paraphrasing process in an MT system can be thought of as a special case of an inter-lingual system in which the inter-lingual is a natural language. Of course, both paraphrasing and inter-lingual systems aim at assisting the transfer by changing input sentences into expressions suitable for transfer engine. The paraphrasing used in our system differs from that in an inter-lingual system, though, in three ways. First, an inter-lingual system requires a symbol system that can cover all possible meanings. Designing such a system is a very complicated task. Paraphrasing does not need such a system because any meaning can be expressed in natural language. Second, the inter-lingual system and the transfer engine depend on each other; therefore, development or improvement of either will affect the other. In contrast, our paraphraser and transfer engine can be developed independently because the natural language used between them can be accepted by either side. Third, inter-lingual expressions are difficult to understand. Paraphrasing results are expressed in natural language, though, and thus are easily understood by the developers of both the paraphraser and the transfer engine. In addition, natural language expressions can be easily retrieved from a database, so paraphrasing results are more useful for an example-based MT system.

### *2.2 Goals for paraphrasing*

We have three goals for the paraphrasing process (Zhang and Yamamoto 2001; Zong, Zhang, Yamamoto, Sakamoto and Shirai 2001).

**Goal I:** Transform ungrammatical expressions in spoken language into formal expressions. For example, in spontaneous utterances, words are often uttered in an order very different from that in written communication. Unscrambling the inverted word orders helps the transfer process.

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

**Goal II:** Transform complicated expressions into simple ones. Simple expressions have fewer ambiguities, so they are easier for the transfer engine to translate.

**Goal III:** Generate as many different expressions as possible. Since the transfer engine and the paraphraser are being developed independently, what kinds of expressions can be translated by the transfer is unknown to the paraphraser. When necessary, the paraphraser should be able to generate another alternative expression.

### *2.3 Paraphrasing approach*

We have adopted a pattern-based approach to paraphrasing that utilizes paraphrase instances. For a pair of paraphrase instances, one the original sentence and the other a variation of the sentence, we can use the pair as a paraphrase pattern by taking the original sentence as the part used for matching and the variant as the part to be generated. If the input sentence matches the original sentence, it can be changed to the variant without changing its meaning. However, the coverage is low with such formed patterns because the input sentence seldom matches an original sentence. Paraphrasing thus becomes the task of collecting paraphrase instances, generalizing the instances into more abstract patterns, and applying patterns to the input sentences. We will describe these tasks in the following sections.

The pattern-based approach requires only that the sentences be part of speech (POS) tagged; it does not require deep analysis. This approach is therefore practical for Chinese paraphrasing as techniques for parsing and semantic analysis are not well-developed enough for practical use in Chinese. In the case of spoken language especially, the problems of parsing and semantic analysis are themselves unresolved.

If the transfer engine fails to translate a sentence, it specifies to the paraphraser which part cannot be processed and asks it to provide an alternate sentence. With this information, the paraphraser identifies paraphrase patterns that vary the specified part. In this way, the paraphraser can effectively change the sentence into one that is more likely to be translated by the transfer engine. For this purpose, we classify paraphrase phenomena based on different aspects of language phenomena and acquire corresponding paraphrase patterns. In the paraphrasing process, the paraphraser can apply different patterns depending on the situation. For each of the three paraphrasing goals above, we can select related paraphrase phenomena to process. For example, for Goal I, we can select word order inversion and rearrange the disordered words into a formal word order. Thus, for each goal, we can build a set of paraphrase patterns to be applied.

## 3 Acquisition of paraphrase instances

We collected paraphrase instances from ATR Chinese Paraphrase Corpus.

### *3.1 ATR Chinese Paraphrase Corpus*

Since very little research has been done on Chinese paraphrasing, we started with no knowledge of what expressions can be paraphrased and what expressions can

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

be the result of paraphrasing. We thus required a paraphrase corpus. Although paraphrasing actions often occur in daily communication, paraphrase data have rarely been collected. The ATR Spoken Language Translation Research Laboratory developed a Chinese paraphrase corpus (Zhang, Zong, Yamamoto and Sakamoto 2001) based on the ATR Chinese Travel Conversation Corpus. The original data came from transcripts of a conversation about travel between two speakers, one speaking Japanese and the other speaking English (Takezawa 1999). Since the speakers were allowed to make spontaneous utterances, many spoken language phenomena appeared in the transcripts. The Japanese transcripts were then manually translated into Chinese to produce the ATR Chinese Travel Conversation Corpus. There are 20,000 Chinese sentences in this corpus. These 20,000 sentences were then rephrased manually by native Chinese speakers based on the following principles:

- Rephrasing must keep the meaning the same as that of the original sentence and reflect the characteristics of spoken Chinese as much as possible.
- The words can be reordered, some words can be replaced with synonyms, and the syntactic structure can be changed.
- For each original sentence, at least two different sentences must be given.
- Each original sentence must be rephrased by two people separately. The obtained sentences and the original sentence must be examined by a third person. From these sentences, the third person selected one formal expression and the two expressions that were the closest in meaning to the spoken sentence.

The constructed paraphrase corpus contains 20,000 original sentences and 44,480 variants. There are at least three expressions for each meaning, among which at least one is a formal expression and the others are spoken language expressions or formal expressions. The average length of the 64,480 sentences is 11.7 Chinese characters. Since the content of the corpus is travel conversation, about 49% of the sentences are questions. We give examples of the paraphrase corpus in the following section.

### *3.1.1 Examples of paraphrasing at the lexical level*

[3 — 1 — 1]
**Original:** 有带 浴室 的房间吗?
(you3 dai4 yu4shi4 de fang2jian1 ma?)
'Are there rooms with a bathroom?'
**Variant:** 有带 洗澡间 的房间吗?
(you3 dai4 xi3zao3jian1 de fang2jian1 ma?)
The word "浴室" 'bathroom' and "洗澡间" 'bathroom' can be substituted for each other in most cases.
[3 — 1 — 2]
**Original:** 房间里有 电视机 吗?
(fang2jian1 li3 you3 dian4shi4ji1 ma?)
'Is there a TV set?'

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

**Variant:** 房间里有 电视 吗?
(fang2jian1 li3 you3 dian4shi4 ma?)
'Is there a TV?'

The word "电视" has two meanings – one is "TV set", the other is "TV program". In this context, "电视机" and "电视" have the same meaning: "TV set".

### *3.1.2 Example of paraphrasing at the phrase level*

[3 – 1 – 3]
**Original:** 这个房间 太贵了 。
(zhe4ge fang2jian1 tai4gui4le.)
'This room is too expensive.'
**Variant:** 这个房间 价钱太高了 。
(zhe4ge fang2jian1 jia4qian2 tai4gao1le.)
'The price of this room is too high.'
The phrases "太贵了" 'too expensive' and "价钱太高了" 'The price is too high' have basically the same meaning and can be substituted for each other in most cases.

### *3.1.3 Examples of paraphrasing at the syntax level*

[3 – 1 – 4]
**Original:** 请给我 发票 。
(qing3 gei3 wo3 fa1piao4.)
'Please give me a receipt.'
**Variant:** 请 把发票 给我。
(qing3 ba3 fa1piao4 gei3 wo3.)
'Please give a receipt to me.'
[3 – 1 – 5]
**Original:** 用金卡付还是用现金 付?
(yong4 jin1ka3 fu4 hai2shi4 yong4 xian4jin1 fu4?)
'Will you pay by card or by cash?'
**Variant:** 金卡和现金 用哪个付?
(jin1ka3 he2 xian4jin1 yong4 na3ge fu4?)
'How would you prefer to pay, card or cash?'

## *3.2 Automatic extraction of instances*

By pairing each formal expression with its corresponding variants, we obtained 44,480 "paraphrase pairs". Word segmentation and part of speech tagging were then carried out. The part of speech tagger used the Penn Chinese Treebank tag set, which has 33 categories (Xia 2000).

To acquire patterns related to a certain paraphrase phenomenon, we first collect the paraphrase instances containing that paraphrase phenomenon. We select one paraphrase phenomenon for each of the three paraphrasing goals. Here, we use them to illustrate how we collect the related paraphrase instances.

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

### *3.2.1 Word order inversion*

In written Chinese, the formal word order approximately follows the SVO (subject-verb-object) pattern. Additionally, time representations and locality representations usually come before the verb, and the modifier usually comes before the part modified. In spoken Chinese, in contrast, word order is very free. As part of achieving Goal I, we unscramble the inverted words into a formal word order (Zhang, Zong, Yamamoto and Sakamoto 2001). From the paraphrase corpus, we automatically extract paraphrase pairs in which the numbers of morphemes in the two sentences are the same and each morpheme in one sentence appears in the other one and vice versa. In the extracted paraphrase pairs, we found the following types of word order inversion.

(1) Prepositional phrases
[3 – 2 – 1]
**Formal Expression:** 对房间 有什么要求吗?
(dui4 fang2jian1 you3 shen2me yao1qiu2 ma?)
'Are there any requests for rooms ?'
**Spoken Expression:** 有什么要求吗? 对房间.
(you3 shen2me yao1qiu2 ma? dui4 fang2jian1)

In the formal expression, the prepositional phrase "**对房间**" 'for rooms' comes at the beginning of the sentence. In the spoken expression, it comes at the end of the utterance.

(2) Auxiliary verbs
[3 – 2 – 2]
**Formal Expression:** 我应该 告诉谁好呢?
(wo3 ying1gai1 gao4su4 shei2 hao3 ne?)
'Whom should I tell?'
**Spoken Expression:** 告诉谁好呢? 我应该
(gao4su4 shei2 hao3 na, wo3 ying1gai1?)

(3) Objects of verbs
[3 – 2 – 3]
**Formal Expression:** 能告诉我 您的名字 吗?
(neng2 gao4su4 wo3 nin2 de ming2zi ma?)
'May I have your name?'
**Spoken Expression:** 您的名字 能告诉我吗?
(nin2 de ming2zi neng2 gao4su4 wo3 ma?)

(4) Modifiers and the modified words
[3 – 2 – 4]
**Formal Expression:** 没有 量少一点儿的 药 吗?
(mei2you3 liang4 shao3 yi1dianr3 de yao4 ma?)
'Is there medicine in small volume?'
**Spoken Expression:** 药 没有 量少一点儿的 吗?
(yao4 mei2you3 liang4 shao3 yi1dianr3 de ma?)

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

(5) Interrogative pronouns
[3 – 2 – 5]
**Formal Expression:** 几点 到上海呢?
(ji2dian3 dao4 shang4hai3 ne?)
'What time will we arrive in Shanghai?'
**Spoken Expression:** 到上海 几点 呢?
(dao4 shang4hai3 ji2dian3 ne?)

### 3.2.2 Negative expressions

Recognition of negative information in a sentence is important and usually requires special processing. In the paraphrase corpus, we found some cases where the original sentences included negative words "不" 'do not' or "没" 'did not' while the corresponding variants did not. This indicates that some sentences with negative expressions can be simplified into affirmative expressions. As part of achieving Goal II, we try to simplify the sentences with negative expressions into affirmative expressions. To acquire this negative/affirmative paraphrase patterns, we automatically extracted paraphrase pairs in which one sentence included the words "不" or "没" and the other did not. Examples are shown in [3 – 2 – 6] and [3 – 2 – 7].

[3 – 2 – 6]
**Original:** 不到 中午 不能 登记。
(bu2 dao4 zhong1wu3 bu4neng2 deng1ji4.)
'You cannot check in until noon.'
**Variant:** 到了 中午 才能 登记。
(dao4 le zhong1wu3 cai2neng2 deng1ji4.)
'you can check in only when it is noon.'

[3 – 2 – 7]
**Original:** 知不知道 我的電話?
(zhi1 bu4 zhi1dao4 wo3 de dian4hua4?)
'Don't you know my telephone number?'
**Variant:** 知道 我的電話吗?
(zhi1dao4 wo3 de dian4hua4 ma?)
'Do you know my telephone number?'

### 3.2.3 Structure of "把" 'ba'

To generate as many different expressions as possible, we select the "把" 'ba' structure. As part of achieving Goal III, we can enrich an expression by inserting or deleting the "把" structure. An expression with the pattern "S(subject) V(verb) O(object) C(complement)" can be changed into a "S 把 O V C" expression by inserting "把", or the opposite can be done by deleting "把" (Zhang and Sato 1999). The two structures have basically the same meaning. We automatically extracted paraphrase pairs in which one sentence included the particle "把" and the other did not. An example is given in [3 – 2 – 8].

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

[3 — 2 — 8]
**Original:** 这张单子 请您填好.
(zhe4 zhang1 dan1zi3 qing3 nin2 tian2hao3)
'Could you fill out this form, please.'
**Variant:** 请您 把这张单子 填好.
(qing3 nin2 ba3 zhe4 zhang1 dan1zi3 tian2hao3)
'Could you make this form filled out, please.'

We automatically extracted three kinds of paraphrase pairs related to these three cases. Because one paraphrase pair may include more than one of these cases, the same paraphrase pair may be extracted more than once for different cases.

## 4 Generalization of instances into a pattern

We generalized the extracted paraphrase pairs into more abstract expressions, called "paraphrase patterns", to match more sentences. As mentioned above, each paraphrase pattern consists of two parts, the part used for matching and the generated part. For each paraphrase pair, if the paraphrase pair was related to Goal I, the spoken language sentence was generalized into the matching part, and the formal expression was generalized into the generated part; if the paraphrase pair was related to Goal II, the sentence of the complicated expression was generalized into the matching part, and that of the simple expression was generalized into the generated part; and if the paraphrase pair was related to Goal III, one of the two sentences was generalized into the matching part, and the other was generalized into the generated part.

### *4.1 Automatic generalization*

In generalization, we disregard the restrictions of some orthographic expressions or parts of speech by replacing them with variables. A generalized pattern is shown in [4 — 1 — 1].

[4 — 1 — 1] 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

The matching part is to the left of the arrow; it defines the constituents to be paraphrased and the restrictions on context. $X_i$ in the matching part can match any orthographic expression in the matching process. The generated part is to the right of the arrow; it defines how to rearrange the constituents to form a new sentence. $X_i$ in the generated part defines a substitution operation in sentence generation.

This pattern is obtained by generalizing the paraphrase pair in [4 — 1 — 2]. Each morpheme has two attributes, orthographic expression and parts of speech, separated by "/".

[4 — 1 — 2] A paraphrase pair.
这/DT 张/M 单子/NN 请/VV 您/PN 填好/VV→
请/VV 您/PN 把/BA 这/DT 张/M 单子/NN 填好/VV

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

In the generalization of instances, we face a trade-off between generalization and accuracy. Disregarding some restrictions in application can make patterns more general while loosened restrictions can result in incorrect results. In our paraphrasing task, the generated sentences must be correct Chinese sentences and must also have the same meanings as the input sentences. Therefore, we have to consider accuracy first.

Since parts of speech are important in determining the syntactic structure, we reserved all parts of speech. Verbs, auxiliary verbs, adjectives, adverbs, and some function words basically determine the syntax of a sentence and are important in retaining the meaning. We therefore reserved the orthographic expressions of these parts of speech. For the other parts of speech, nouns, pronouns, numerals, etc., we replaced their orthographic expressions with variables. [4 — 1 — 1] is obtained by generalizing [4 — 1 — 2] in this way.

A preliminary experiment revealed two problems with this automatic generalization. First, a matching part generalized from a long sentence was often too restricted to match other sentences because too many attributes were reserved as restrictions. In fact, some clauses and noun phrases that functioned as modifiers were not referred to in the paraphrase phenomenon under consideration, and therefore their attributes did not have to be reserved as restrictions. We can thus disregard such clauses and noun phrases. Second, some orthographic expressions categorized as nouns, pronouns, or numerals were automatically replaced with variables, although they played key roles in determining the meaning of the sentence. For example, the numeral "多少" 'how many' is different from other numerals, such as "2003", because it may make the sentence interrogative. Such orthographic expressions should therefore be reserved while other orthographic expressions in the same categories can be disregarded. To resolve these two problems, we need human intervention to complement the automation of generalization.

### *4.2 Manually assisted generalization*

The above automatic generalization process needs to be told which orthographic expressions can be disregarded and which ones should be reserved. We have developed an editing tool that enables people to easily specify such information and generalize the defined instances into patterns (Zhang, Zong, Yamamoto and Sakamoto 2001). The tool consists of description symbols and a transformation program. The symbols are used to specify the constituents of the instances, and the transformation program automatically transforms the edited instances into patterns. Two description symbols are defined as follows.

**[ ]:** This symbol is followed by a number and is used to enclose a sequence of morphemes in both the original sentence and the variant. The enclosed part is a syntactic constituent, e.g. a noun phrase or a clause. In transformation, only the part of speech of the last morpheme is reserved; the other attributes are replaced with one variable. Since the syntactic role of a sequence of words in the Chinese language is likely to be reflected in the last word, the part of speech of the last morpheme is reserved. The enclosed part in the original sentence and

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

that in the variant denoted by the same number are replaced with the same variable.

{ }: This symbol is used to enclose a morpheme. In transformation, the orthographic expression of the morpheme is reserved to prevent it being replaced with a variable. In this way, the orthographic expression of the morphemes can be reserved as a restriction. Two or more orthographic expressions can be defined inside this symbol so that words that can be paraphrased in the same way can be stored in one pattern.

The transformation program automatically generalizes the edited instances into patterns as follows.

- Replace the parts enclosed by the symbol [ ] or { } with variables following the above definition.
- Replace the orthographic expressions of nouns, pronouns, and numerals with variables.

A generalized pattern obtained using this tool is shown in [4 − 2 − 1].

[4 − 2 − 1] The generalized pattern.

$Y_1$/DEG $Y_2$/NN 请/VV 给/VV $X_1$/PN $X_2$/CD 本/M →
请/VV 给/VV $X_1$/PN $X_2$/CD 本/M $Y_1$/DEG $Y_2$/NN

In [4 − 2 − 1], $X_i$ has the same meaning as in [4 − 1 − 1]. $Y_1$/DEG in the matching part means that it can match the sequence of morphemes in which the part of speech of the last morpheme is equal to DEG. $Y_1$/DEG in the generated part defines a substitution operation for sentence generation. $Y_2$/NN means the same thing, but the part of speech of the last morpheme must be NN.

The use of the tool is illustrated as follows. We edit the paraphrase pair manually using description symbols, as shown in [4 − 2 − 2]. The edited instance is then automatically transformed into the pattern [4 − 2 − 1] by the transformation program.

[4 − 2 − 2] The edited paraphrase pair.

[日语/NN 的/DEG]$_1$ [指南/NN 手册/NN]$_2$ 请/VV 给/VV 我/PN 两/CD 本/M→

(ri4yu3 de zhi3nan2 shou3ce4 qing3 gei3 wo3 liang3 ben3)

'Could you give me two copies of the Japanese pamphlet, please?'

请/VV 给/VV 我/PN 两/CD 本/M [日语/NN 的/DEG]$_1$ [导游/NN 手册/NN]$_2$

(qing3 gei3 wo3 liang3 ben3 ri4yu3 de dao3you2 shou3ce4)

'Could you give me two copies of the Japanese tourist guide, please?'

This paraphrase pair includes two paraphrase phenomena: one is word order inversion, and the other is the substitution of the synonymous noun phrases "指南 'guide' 手册 'handbook'" and "导游 'tourist guide' 手册". Since this example is related to the pattern of word order inversion, the synonymous noun phrases "指南 手册" and "导游 手册" are enclosed by [ ] with the same number, 2. As a result, both are generalized to $Y_2$/NN. In this way, the substitution of synonymous noun phrases is disregarded, and only word order inversion is considered.

Another example of the generalized pattern is shown in [4 − 2 − 3]. The edited paraphrase pair is shown in [4 − 2 − 4].

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

[4 – 2 – 3] Another example of the generalized pattern.
$X_1$/NR $X_2$/PN (想|打算|准备)/VV 去/VV 看看/VV →
$X_2$/PN (想|打算|准备)/VV 去/VV 看看/VV $X_1$/NR

[4 – 2 – 4] The edited paraphrase pair.
南大门/NR 我/PN {想, 打算, 准备/VV} 去/VV 看看/VV →
(nan2da4men2 wo3 {xiang3, da3suan4, zhun3bei4} qu4 kan4kan4.)
'I {want, plan, prepare} to go to see Namdaemun.'
我/PN {想, 打算, 准备/VV} 去/VV 看看/VV 南大门/NR

The original sentence of the paraphrase pair is "南大门/NR 我/PN 想/VV 去/VV 看看/VV", and the variant is "我/PN 想/VV 去/VV 看看/VV 南大门/NR". This is a paraphrase pair related to word order inversion. Other verbs, for instance, "打算" 'plan' and "准备" 'prepare', whose meanings are near to the main verb "想" 'want', can be reordered in the same way. We added the verbs "打算" 'plan' and "准备" 'prepare' into the symbol { }. Thus, the obtained pattern [4 – 2 – 3] can cover more cases. In this way, the coverage of the obtained patterns can be increased without loss of accuracy.

### *4.3 Construction of the paraphrase pattern*

Using the two description symbols, we manually edited the 1,634 paraphrase pairs extracted as described in section 3.2. This manual editing work required about one and a half months for one person to finish. The edited instances were then generalized into patterns by the transformation program. The obtained paraphrase patterns were categorized into four groups.

$G_{WordOrder}$: 1,015 patterns for reordering words from spoken-language style into formal word order.

$G_{NegativeExpression}$: 459 patterns for simplifying negative expressions into affirmative expressions.

$G_{InsertBa}$: 160 patterns for inserting "把" 'ba'.

$G_{DeleteBa}$: 160 patterns for deleting "把" 'ba'.

## 5 Implementation of the paraphraser

The paraphrasing process is actually a process of applying patterns to input sentences. The application order is as follows (Zhang and Yamamoto 2002). Patterns aimed at achieving Goal I are applied first, then those patterns aimed at Goal II and finally those patterns aimed at Goal III are applied. We number the groups of patterns aimed at Goal I from 1, then continue the numbering with the groups aimed at Goal II and lastly with the groups aimed at Goal III. Let the last number be $N$. Thus, the groups of patterns are expressed as $G_1, G_2, \ldots, G_N$. For example, we index the pattern groups obtained above $G_{WordOrder}$, $G_{NegativeExpression}$, $G_{InsertBA}$, and $G_{DeleteBA}$ as $G_1, G_2, G_3$, and $G_4$, respectively.

For the different groups aiming at the same goal and the different patterns in the same group, we determine their application order randomly. To generate as many

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

different expressions as possible, the sentences generated by the preceding groups are also processed by all subsequent groups. In this application mechanism, the spoken language expressions are first transformed into formal expressions, then the complicated structures are simplified, and finally the expressions are enriched. Since the spoken language and complicated expressions are transformed into formal and simple expressions in the early phase, they should be correctly processed in the following paraphrasing. The paraphrasing process is as follows.

(1) Put the input sentence into the data stack, $D_i$, of each pattern group $G_i$, $1 \leq i \leq N$. Set the pointer of pattern group $j$ equal to 1.
(2) Apply each pattern of $G_j$ to each sentence of $D_j$.
    (2.1) Match the sentence with the matching part of the pattern. If the matching succeeds, generate a sentence using the generated part of the pattern.
    (2.2) Put the generated sentence into $D_{j+1}, \ldots, D_N$.
(3) Set $j$ equal to $j+1$, and go to step (2) until $j > N$.
(4) Filter out duplicate sentences when outputting the generated sentences.

There are two notable characteristics of this process. First, the generated sentences can be passed to the transfer engine as soon as they are produced. Once a generated sentence is translated by the transfer engine, the paraphraser can stop the paraphrasing process. Second, since the generated sentences are further processed using the patterns of all subsequent groups, more expressions are likely to be generated.

## 6 Experiment and evaluation

Since our objective is to increase the robustness of the SLT system with regard to spoken language, the performance of the paraphraser should be evaluated in terms of the improvements the paraphraser brings to the SLT system. However, integration of the implemented paraphraser with other modules – the speech recognizer and the transfer engine – will not be possible in the immediate future because the other modules are still being developed, so we cannot do such an evaluation at present. Therefore, our preliminary evaluation of the paraphraser was done with respect to, for example, whether the generated sentences are correct Chinese sentences, whether the meanings of the input sentences are retained by the generated sentences, and how many different expressions are generated for each input sentence.

### *6.1 Experiment*

The implemented paraphraser and obtained paraphrase patterns were evaluated through a paraphrasing experiment. The paraphrase patterns were $G_{WordOrder}$, $G_{NegativeExpression}$, $G_{InsertBA}$, and $G_{DeleteBA}$. The test data were 45,110 simple sentences from the paraphrase corpus. To achieve the same results as when using open test data, when a sentence was processed, we did not apply the patterns generalized from that sentence.

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

Table 1. *Evaluation result*

| | |
|---|---|
| No. Test sentences | 45,110 |
| No. Test sentences matched paraphrase pattern | 4,908 (10.9%) |
| No. Generated sentences | 8,183 |
| No. Correct paraphrases | 7,226 (88%) |

### *6.2 Evaluation*

Application of the paraphrase patterns to the 45,110 test sentences resulted in 4,908 test sentences matching one or more of the paraphrase patterns, i.e. the coverage of the paraphrase patterns was about 10.9%. From the 4,908 test sentences, 8,183 sentences were generated. The average number of generated sentences for a test sentence was 1.66. The generated sentences were evaluated by two native speakers of Chinese from two viewpoints: naturalness and meaning retention. The judges evaluated naturalness by determining whether the generated sentence was a correct Chinese sentence. That is, could the judges understand its meaning? Then the corresponding test sentence was presented to the judges. The judges determined whether the understood meaning was the same as that of the presented sentence. The generated sentence was regarded as a correct paraphrase only when it was correct regarding both naturalness and meaning retention. Otherwise, it was considered incorrect. As a result, 7,226 generated sentences were judged as correct paraphrases (accuracy of about 88%). The evaluation results are summarized in Table 1.

### *6.3 Examples of paraphrasing results*

Examples of correct generated paraphrases are shown below.

[6 – 1]

**Test Sentence:** 饭店 在这儿订 可以 吗
(fan4dian4 zai4 zhe4er ding4 ke1yi3 ma)
'a restaurant can I reserve here?'

**Result 1:** 在这儿订 饭店 可以 吗
(zai4 zhe4er ding4 fan4dian4 ke1yi3 ma)
'I want to reserve a restaurant here, can I?'

**Result 2:** 在这儿 可以 订 饭店 吗
(zai4 zhe4er ke1yi3 ding4 fan4dian4 ma)
'Is it possible to reserve a restaurant here?'

**Result 3:** 可以 在这儿订 饭店 吗
(ke1yi3 zai4 zhe4er ding4 fan4dian4 ma)
'Can I reserve a restaurant here?'

The test sentence was spoken language in which the verb "可以" 'can' and its object "饭店" 'a restaurant' were disordered, i.e. the object "饭店" was uttered before the verb "可以". After application of the pattern $G_{WordOrder}$, three new sentences were generated in which the disordered word orders were unscrambled. The three generated sentences were all correct paraphrases.

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

[6 – 2]
**Test Sentence:** 能不能 帮我订到最早的班机呢?
(neng2 bu neng2 bang1 wo3 ding4 dao4 zui4zao3 de ban1ji1 ne?)
'Could you reserve the earliest flight for me?'
**Result 1:** 能 帮我订到最早的班机吗
(neng2 bang1 wo3 ding4 dao4 zui4zao3 de ban1ji1 ma?)
'Could you reserve the earliest flight for me?'
**Result 2:** 可以 帮我订到最早的班机吗
(ke1yi3 bang1 wo3 ding4 dao4 zui4zao3 de ban1ji1 ma?)
'Could you reserve the earliest flight for me?'

The test sentence contained the expression of interrogation, "能不能", which consists of an affirmation "能" 'can' and a negation "不能" 'cannot'. After the application of the pattern for simplifying negative expressions, two new sentences were generated, which appeared as affirmative forms. The two sentences were both correct paraphrases.

[6 – 3]
**Input Sentence:** 风景漂亮的房间 请给我.
(feng1jing1 piao1liang4 de fang2jian1 qing3 gei3 wo3.)
'A room with a nice view, please give me.'
**Result 1:** 请给我 风景漂亮的房间.
(qing3 gei3 wo3 feng1jing1 piao1liang4 de fang2jian1.'
'Please give me a room with a nice view.'
**Result 2:** 请 把 风景漂亮的房间给我.
(qing3 ba3 feng1jing1 piao1liang4 de fang2jian1 gei3 wo3.)
'Please arrange a room with a nice view for me.'
**Result 3:** 我 想要 风景漂亮的房间.
(wo3 xiang3 yao4 feng1jing1 piao1liang4 de fang2jiang1.)
'I would like a room with a nice view.'

In this test sentence, the predicate "请给我" 'please give me' and the object "风景漂亮的房间" 'a room with a nice view' were disordered. After application of the pattern $G_{WordOrder}$, Result 1 was generated in which the disordered word order was unscrambled. The patterns $G_{InsertBA}$ and $G_{DeleteBA}$ were applied to Result 1, and Results 2 and 3 were generated. The three results were correct paraphrases.

Note the following with regard to the meaning-retention criteria. From the viewpoint of linguistics, there are subtle differences, such as differences in nuance, between the test sentences and their results, which were judged to be correct paraphrases. From the viewpoint of communication, any speech action or sentence has one objective that the speaker intends to achieve. If the objective can be equally achieved by another expression, we regard the expression to be a paraphrase of the sentence. The above evaluation carried out by the two judges was based on this consideration. The task thus is domain dependent.

The experimental results show that the proposed approach can simplify the input sentences and produce different expressions with 88% accuracy. If one of the

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

generated expressions can be translated by the transfer engine, we can say that the paraphraser is an effective way to overcome the problem of unrestricted expressions in an SLT system.

### *6.4 Obtained paraphrase patterns*

The results also show that the proposed method actually obtains paraphrase patterns that are related to various paraphrase phenomena. For example, in the paraphrase patterns for reordering words, we found the patterns for reordering the following combinations of two constituents.

- prepositional phrases and main verbs
- interrogative pronouns and main verbs
- some adverbs and main verbs
- verbs and their objects
- auxiliary verb "可以" 'may' and the remaining part of the sentence
- modifiers and modified words
- measure words and modified nouns.

### *6.5 Error analysis*

We also analyzed the incorrect results and found two reasons for errors occurring. One was that some constituents or some modify-modified relations were incorrectly recognized based on the obtained paraphrase patterns. For example, the sentence "请给我两个人住的房间" 'A room for two people, please.' has the syntactic structure "请给我 ((((两个)人)住的)房间)", in which the quantity phrase "两个" 'two' modifies "人" 'people'. When it was matched with pattern [4 — 2 — 1], the quantity phrase "两个" 'two' was incorrectly recognized as modifying "人住的房间" 'a room where people live', i.e. it was recognized as having the syntactic structure "请给我 ((两个)(人住的房间))". Because of this incorrect recognition, the generated sentence was "人住的房间请给我两个" 'Please give me two rooms where people live.' This result was not correct because its meaning differed from the test sentence. This result revealed that shallow parsing is required. The second source of errors was that there were errors in the part of speech tagger we used. For example, the preposition "给" 'for' in the test sentence "请 给 我包装一下这礼物" 'Please gift-wrap this for me' was wrongly tagged as a verb. The wrongly tagged sentences matched unsuitable patterns, resulting in incorrect results.

## 7 Conclusion

We have described the design and implementation of a Chinese paraphraser for a Chinese-Japanese spoken-language translation system. We adopted a pattern-based approach to paraphrasing and developed a method of acquiring paraphrase patterns from a paraphrase corpus, including the classification of paraphrase phenomena, automatic extraction of paraphrase instances related to a paraphrase phenomenon, and generalization of paraphrase instances into paraphrase patterns

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

by integrating human experience. We implemented a paraphraser and obtained four types of paraphrase patterns. The effectiveness of the paraphraser has been tested experimentally. This approach has the following characteristics:

- The pattern-based paraphraser is easy to implement, and the processing time is short because only morphological analysis is required.
- Use of the developed editing tool enables paraphrase patterns to be efficiently learned from a paraphrase corpus and human experience. The coverage of patterns can be increased without loss of accuracy. The patterns augmented by human experience achieved an accuracy of 88% in our paraphrasing experiment.
- The idea of classifying paraphrase phenomena leads to the classification of paraphrase patterns. The paraphraser can paraphrase input sentences with respect to certain aspects of language phenomena by using different types of patterns.
- In the paraphrasing process, the generated sentences are further processed using the other pattern groups. As a result, more alternative expressions are likely to be produced.

In this paper, we have considered three types of paraphrase phenomena and have used them to explain the proposed method. The coverage of the obtained patterns is still low – only 10.9% of the test sentences matched the pattern. Our work is still in the early stages, and the results presented here are only preliminary. We will next investigate more paraphrase phenomena and work on unresolved issues, such as developing a theoretical framework, determining how the paraphrasing process relates to other aspects of natural language processing, and ways to acquire paraphrase knowledge. In particular, we will evaluate the improvement in performance of a Chinese-Japanese SLT system when our Chinese paraphraser is applied. In addition to its application to machine translation systems and information retrieval, our paraphraser can also be applied to second language instruction of Chinese.

## Acknowledgements

This work was done when the authors were at ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. This research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled “A study of speech dialogue translation technology based on a large corpus”. We would like to thank Ms. Lan Yao for her help with our experiments.

## References

Barzilay, R. and Lee, L. (2002) Bootstrapping lexical choice via multiple-sequence alignment. *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 164–171.

Barzilay, R. and McKeown, K. R. (2001) Extracting paraphrases from a parallel corpus. *Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL ’01)*, pp. 50–57.

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press

Dras, M. (1997) Representing paraphrasing using synchronous TAGs. *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97)*, pp. 516–518.

Iordanskaja, L., Kittredge, R. and Polguere, A. (1991) Natural Language Generation in Artificial Intelligence and Computational Linguistics. Kluwer.

Kataoka, A., Masuyama, S. and Yamamoto, K. (1999) Summarization by shortening a Japanese noun modifier into expression "A no B". *Proc. of the 5th Natural Language Processing Pacific-Rim Symposium (NLPRS '99)*, pp. 409–414.

Pang, B., Knight, K. and Marcu, D. (2003) Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. *Proc. of the Human Language Technology and North American Association for Computational Linguistics Conference* (*HLT/NAACL*).

Ramsay, A. and Seville, H. (2000) Unscrambling English word order. *Proc. of the 18th International Conference on Computational Linguistics (Coling '00)*, pp. 663–669.

Sato, S. (1999) Automatic paraphrase of technical papers. *Trans. Infor. Process. Soc. Japan*, pp. 2937–2945.

Shirai, S., Ikehara, S. and Kawaoka, T. (1993) Effects of automatic rewriting of source language within a Japanese to English MT system. *Proc. of 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '93)*, pp. 226–239.

Stede, M. (1996) Lexical paraphrases in multilingual sentence generation. *Machine Translation*, **11**: 75–107.

Takezawa, T. (1999) Building a bilingual travel conversation database for speech translation research. *Proc. of Oriental COCOSDA Workshop*, pp. 17–20.

Xia, F. (2000) The Part-of-speech Tagging Guideline for the Penn Chinese Treebank (3.0). *http://www.ldc.upenn.edu/ctb*.

Yamamoto, K., Shirai, S., Sakamoto, M. and Zhang, Y. (2001) Sandglass: twin paraphrasing spoken language translation. *Proc. of the 6th International Conference on Computer Processing of Oriental Languages (ICCPOL '01)*, pp. 154–159.

Yamamoto, K. (2002) Machine translation by interaction between paraphraser and transfer. *Proc. of the 19th International Conference on Computational Linguistics (Coling 2002)*, pp. 1107–1113.

Zhang, L. and Sato, H. (1999) Chinese Expression Grammar-28 Points. Toho-Shoten.

Zhang, Y. and Yamamoto, K. (2001) Analysis of Chinese spoken language for automatic paraphrasing. *Proc. of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL '01)*, pp. 290–293.

Zhang, Y., Zong, C., Yamamoto, K. and Sakamoto, M. (2001) Paraphrasing utterances by reordering words using semi-automatically acquired patterns. *Proc. of the 6th Natural Language Processing Pacific-Rim Symposium (NLPRS '01)*, pp. 195–202.

Zhang, Y. and Yamamoto, K. (2002) Paraphrasing of Chinese utterances. *Proc. of the 19th International Conference on Computational Linguistics (Coling '02)*, pp. 1163–1169.

Zong, C., Zhang, Y., Yamamoto, K., Sakamoto, M. and Shirai, S. (2001) Approach to spoken Chinese paraphrasing based on feature extraction. *Proc. of the 6th Natural Language Processing Pacific-Rim Symposium (NLPRS '01)*, pp. 551–556.

Zukerman, . and Raskutti, B. (2002) Lexical query paraphrasing for document retrieval. *Proc. of the 19th International Conference on Computational Linguistics (Coling '02)*, pp. 1177–1183.

https://doi.org/10.1017/S1351324905003712 Published online by Cambridge University Press