ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

ANALYSIS OF PARAPHRASED CORPUS AND LEXICAL-BASED APPROACH TO CHINESE PARAPHRASING

Yan ZHANG and Hideki KASHIOKA

ATR Spoken Language Translation Research Laboratories
2-2-2 Keihanna Science City, Kyoto 619-0288
{yan.zhang, hideki.kashioka}@atr.jp

ABSTRACT

In this paper, we firstly analyze the language phenomena and distribution characteristics of Chinese spontaneous utterances already paraphrased by other approaches. Based on the information obtained from a corpus, our lexical-based approach is proposed to paraphrase Chinese spoken language. Our purpose is to transform various expressions into simplified expressions with the same meanings. Chinese verbs are the main constituents in sentences, and with their flexibility they play an important role in expressing structures, especially for transitive verbs. Furthermore, negative verb expressions also appear frequently to express enquiries in question utterances. Therefore, we design four types of paraphrasing templates based on lexical information and the characteristics of the corpus: (1) synonym replacement, (2) Chinese transitive verbs, (3) verbs with two objects, and (4) the transformation of negative expressions. Our experiment found that the lexical-based approach is effective for Chinese paraphrasing.

1. INTRODUCTION

In the spoken language translation (SLT) field, a key issue is handling flexibility and unrestricted expressions. Therefore, a wide variety of research on SLT has been conducted, and different approaches have been proposed to handle spoken phenomena over the past decade. Paraphrasing is an appropriate way to deal with this problem and provide more variant expressions. It provides a way to pre-process the spontaneous utterances prior to transfer [1, 2]. Many possible expressions for a particular meaning can be generated due of the flexibility of spontaneous utterances. For a machine translation system, when an expression can't be handled, other expressions are generally available to replace the current one in spoken language.

Paraphrasing approaches have been presented and applied widely in SLT systems, especially in the English [3] and Japanese languages [1, 2, 4]. However, there is a limited number of research reports on Chinese language paraphrasing [5, 6, 7, 8]. These papers provide the main framework by summarizing templates to use in Chinese paraphrasing. More sophisticated method of Chinese language processing would require resources such as syntactically or semantically annotated databases. However, such resources are scarce. Therefore, a lexical-based approach is proposed in line with the characteristics of spontaneous utterances. This approach focuses on four aspects: (1) synonym replacement, (2) Chinese transitive verbs, (3) verbs with two objects, and (4) the transformation of negative expressions.

In order to investigate the phenomena and distribution characteristics of the Chinese paraphrasing corpus, we analyzed the paraphrased results already processed by other approaches. The paraphrased corpus is from the ATR corpus of hotel reservation utterances, including 64,477 Chinese utterances. Each utterance is paraphrased into two utterances on average.

This paper is organized into the following sections. Section 2 analyzes the phenomena of the already paraphrased utterances. Here we analyze the distribution properties and language information with a view to developing our new approach. Section 3 is the main part of this paper, and it explains in detail the paraphrasing process used to deal with Chinese verbs and negative expressions. Section 4 gives the experimental results and analysis. Finally, section 5 draws our conclusions about the proposed lexical-based approach and discusses our future work on Chinese paraphrasing.

2. ANALYSIS OF PARAPHRASED CORPUS

The corpus in this paper consists of expressions used in the hotel reservation domain, where each utterance has been paraphrased. Y.J. Zhang [9] has investigated the phenomena of the Chinese spoken language in the original corpus before paraphrasing. These utterances in the corpus provide important information for learning new structures and patterns for use in the lexical-based approach. Moreover, the corpus has been processed from three layers: word segmentation, part-of-speech tagging with the Penn Chinese Treebank tagset [11], and partial phrase identification. After analyzing the corpus, we found some problems at different layers that affect Chinese paraphrasing and also summarized the distribution characteristics of the utterances.

2.1. Problems in the Corpus

(1) Chinese word segmentation

[ex.1] 能 介绍 我家 不 太 贵 的 饭店 吗 ?

(Could you introduce me to a hotel whose rates are not too expensive?)

In the sentence, the word “我家” should be segmented to “我”and “家” because “家” is a quantifier here.

(2) Part-of-speech tagging

[ex.2] 收/VV 没/AD 收到/VV 日本/RN 佐/X 藤/X 的/DEG 房间/NN 预约/NN ？ /PU

(Have you received Mr. Sato’s room reservation from Japan?)

In above example, tagger ‘X’ means that the Chinese character is the out-of-vocabulary one in the lexicon. This error usually occurs when the Chinese word is the name of a person or place or it has been segmented incorrectly.

(3) Phrase identification

[ex.3] 没有/VE 靠/P [河边/NN 风景/NN]/NP 漂亮/VA 的/DEC 房间/NN 吗/SP ？ /PU

(Do you have a room with good scenery beside the river?)

The noun phrase in brackets ‘[]’ should be separated and constitute a preposition phrase [靠/P 河边/NN]/PP.

Table 1 lists the accuracies for different levels of processing. Generally, these three kinds of errors affect each other and thus further degrade paraphrasing results.

Processing	Accuracy (%)
Word segmentation	96.1
Part-of-speech tagging	93.4
Phrases in chunk identification	70.2

Table 1 Results of corpus preprocessing

2.2. Distribution Characteristics of the Corpus

It is necessary to analyze and summarize the characteristics and distribution properties of the corpus before executing Chinese paraphrasing. Our lexical-based approach uses just such language information. Chinese part-of-speech tags provide fundamental information for Chinese paraphrasing. Chinese verbs are the most important elements in expressing the meanings and structures of Chinese sentences, especially transitive verbs. Therefore, we mainly consider the roles and properties of transitive verbs (denoted by ‘VT’) and verbs with two objects (denoted by verbs + O_1 + O_2, briefly VOO) from the corpus, although VOO actually belongs to VT. Another kind of useful information comes from the negative expressions denoted as ‘A-不-A’ and ‘不-A’, which appear in question utterances used to obtain information.

There are a total of 64,477 Chinese utterances in this hotel reservation corpus. We extracted a Chinese lexicon in which each Chinese word includes three parts: word name, part-of-speech, and occurrence frequency in the corpus. There are 6,801 Chinese words and 1,350 verbs after deleting the affix words and out-of-vocabulary words tagged with ‘X’. Furthermore, the statistical proportions of verbs ‘VT’ and ‘VOO’ in the lexicon are different from those in the corpus. Question utterances occur at a rate of 50% in the entire corpus, and nearly 60% of them contain negative expressions. The analysis results and distribution proportions are displayed in Table 2, in which the symbol ‘--‘ denotes that there is no information in the corresponding item.

Expression or utterance	Proportion (%) in verb lexicon	Proportion (%) in corpus
transitive verbs (VT)	22.07	58.84
verbs + O_1 + O_2 (VOO)	1.48	9.87
Negative expressions in question utterances	--	59.25
Question utterances	--	49.01

Table 2 Distribution proportions of structures

3. LEXICAL-BASED PARAPHRASING APPROACH

Currently there are different levels of approaches to paraphrasing [3, 4, 9], but the lexical-based paraphrasing is considered the first level because it considers the meanings of the words for Chinese paraphrasing. Although this approach is relatively simple as a paraphrasing method, it is valid for Chinese paraphrasing due to the limited Chinese processing techniques as well as the difficulty of obtaining highly accurate Chinese information.

3.1. Synonymous Lexicon

The construction of a synonymous lexicon was carried out to build direct mapping relations among Chinese words. This lexicon mainly includes verbs and nouns extracted from Chinese utterances. The process is performed by first looking up candidates in the synonymous lexicon according to the source word items and then replacing the current words to construct new expressions. Table 3 gives some examples of synonymous words, where paraphrased results are obtained by replacing the synonymous words.

Source words	Synonymous words
想 (VV)	想要，打算，计划

问问(VV)	打听，了解
保存(VV)	保管，保留
供应(VV)	提供，供给
柜台(NN)	服务台，信息台

Table 3 Synonymous lexicon

[ex.4] 推荐/VV 我/PN 带有/VV 洗衣室/NN 的/DEG 饭店/NN 。/PU
(Please recommend me a hotel with a washing room.)
→ 介绍/VV 我/PN 带有/VV 洗衣室/NN 的/DEG 饭店/NN 。/PU

[ex.5] 柜台/NN 里/LC 有/VE 日本人/NN 吗/SP ？/PU
(Is there a Japanese person at the information desk?)
→ 服务台/NN 里/LC 有/VE 日本人/NN 吗/SP ？/PU

In [ex.4] and [ex.5], verbs "推荐" and "介绍" are synonymous verbs, and nouns "柜台" and "服务台" are synonymous in the lexicon, respectively.

3.2. Chinese Verb Structures

We have obtained statistical information on the Chinese verbs in the utterances shown in section 2. Based on the characteristics of the corpus, our Chinese paraphrasing mainly processes two kinds of Chinese verbs: transitive verbs and objective verbs-VOO.

(1) Transitive verbs (VT)
In general, transitive verbs describe a noun or noun phrase as its object. This 'V-O' structure is very flexible, and it can be placed in different positions of a Chinese sentence without changing the meaning of the sentence, especially in spontaneous utterances. For example,

[ex.6] 我/PN 想/VV 换/VV 航班/NN 。/PU
(I want to change the flight.)
→ 换/VV 航班/NN 我/PN 想/VV 。/PU
→ 航班/NN 我/PN 想/VV 换/VV 。/PU

Transitive verb "换" modifies the object "航班". The verb structure can be put at the end of the sentence. And in another form, the object "航班" can be separated from the verb. The two paraphrased sentences are both correct and express the same meaning.

(2) Verbs + O_1 + O_2 (VOO)
The distinctive characteristic of this kind of verb is that it usually describes two objects: a direct object and an indirect object. Furthermore, the direct object can be separated from the VOO structure and put at the head, the end, or another unfixed position in the utterance.

[ex.7] 请/VV 给/VV 我/PN 诊断书/NN 。/PU
(Please give me the certificate of diagnosis.)
→ 诊断书/NN 请/VV 给/VV 我/PN 。/PU

In this example, "我" and "诊断书" are indirect and direct objects, respectively. The position of the direct object changes from the end to the head of the sentence after paraphrasing.

3.3. The Transformation of Negative Expressions

A negative expression, uniformly denoted by 'A-不-A', where 'A' is a verb, is an important structure in question sentences. Actually, the former 'A' and the latter 'A' are not always the same. Sometimes, the latter verb includes the former one. For example, "通/VV 没/AD 通知/VV" and "确/VV 不/AD 确定/VV". There are two transformation directions used to paraphrase a negative structure:

（1）A-不-A → 不-A

The former expression includes two verbs in addition to the negative word "没" or "不". These two expressions have different particles at the end of a sentence. With the change to a negative expression, the sentence-final particle also needs to change. The corresponding relations with particles are described in (A) and (B).

[A]. A-不-A …（呢/SP）？ Particle "呢/SP" can be inserted at the end of the sentence or omitted.

[B]. 不-A ... 吗/SP ？ /PU

[ex.8] 今天/NT 晚上/NT 单人房/NN 有没有/VE 空房/NN ？ /PU
(Do you have a vacant single room tonight?)
→ 单人房/NN 今天/NT 晚上/NT 有/VE 空房/NN 吗/SP ？ /PU

（2）不-A → A-不-A

This transformation is the reverse process of form (1).

[ex.9] 没/AD 收到/VV 日本/NR 佐藤/NN 的/DEG 预约/NN 吗/SP ？ /PU
(Have you received Mr. Sato's reservation from Japan?)
→ 收/VV 没/AD 收到/VV 日本/NR 佐藤/NN 的/DEG 预约/NN 呢/SP ？ /PU
→ 日本/NR 佐藤/NN 的/DEG 预约/NN 收/VV 没/AD 收到/VV 呢/SP ？ /PU

4. EXPERIMENTS AND ANALYSIS

Our Chinese paraphrasing system was performed by randomly selecting about 100 original utterances from the same hotel reservation domain mentioned in section 3. The test utterances have been segmented and tagged with part-of-speech information. In the test set, 40 percent of these utterances are question sentences and the others are statement utterances. The detailed distribution information of utterances and the corresponding paraphrased results are presented in Table 4.

Form of utterance	Property (%) in original utterances	Paraphrased rate (%) of each form
Question utterances	39	50.0
Statement utterances	61	37.7
Synonym replacement	20	35.0
Transitive verbs	31	35.5
VOO	10	40.0
Negative expressions	20	80.0

Table 4 Results of Chinese paraphrasing

In the entire experiment, 41% of the utterances were paraphrased. On average, a Chinese utterance was paraphrased into 1.33 new utterances. After analyzing the paraphrased results, we can summarize the problems and advantages of our approach as follows.

(1) If the utterances are complicated or include multiple clauses, most of the paraphrased utterances are incorrect.
(2) Errors produced by segmentation and part-of-speech tags still constitute a major problem that reduces the proportion of paraphrased utterances.
(3) If the object is a noun phrase or there is a long distance between a transitive verb and its object, the utterance can't be paraphrased or is paraphrased incorrectly. In the row of 'transitive verbs' in Table 4, we can see that the paraphrased rate is only 35.5%.
(4) The coverage of this approach is limited.
(5) Our approach mainly considers key words and some special structures, so the paraphrasing system is easily built and implemented.
(6) The paraphrased rate of negative expressions is comparatively high, at almost 80%, because this structure is relatively fixed.

5. CONCLUSIONS AND FUTURE WORK

Currently, there are limited resources and tools available to perform Chinese paraphrasing. Furthermore, there is not a broad body of research on Chinese paraphrasing, although many paraphrasing techniques can be adapted from other languages. Our preliminary experiment provided a simple but useful idea for developing more effective Chinese paraphrasing for SLT systems.

The approach in this paper, the lexical-based approach, is a low-level approach to Chinese paraphrasing. It has many disadvantages, since it only uses partial language information. On the other hand, it is thus a good way to make use of a Chinese syntactic parser for Chinese paraphrasing. The transformation of sentence structures can provide more important language information than lexical information. Therefore, our future tasks will focus on partial Chinese parsing and full parsing and how to combine Chinese syntactic parsing approaches with Chinese paraphrasing. Furthermore, we will also find a way to evaluate the paraphrased results automatically. The paraphrased results should be simpler and have more regular grammatical sentence structure than the original sentences so that they can improve the performance of machine translation systems.

6. ACKNOWLEDGEMENTS

The research reported here was supported in part by a contract with the Ministry of Public Management, Home Affairs, Posts and Telecommunications entitled "Multilingual speech-translation systems using mobile terminals" and by the National Institute of Information and Communications Technology (NICT) of Japan.

7. REFERENCES

[1] Kazuhide Yamamoto, Satoshi Shirai, Masashi Sakamoto and Yujie Zhang, "Sandglass: Twin Paraphrasing Spoken Language Translation", *Proc. of ICCPOL'01*, Seoul, Korea, pp. 154-159, 2001.

[2] Kazuhide Yamamoto, "Machine Translation by Interaction between Paraphraser and Transfer", *Proc. of COLING2002*, Taipei, pp. 1107-1113, August, 2002.

[3] Andrew Finch, Taro Watanabe and Eiichiro Sumita, "Data-Oriented Paraphrasing", *Proc. of RANLP-2003*, pp.153-157, 2003.

[4] Tokunaga Takenobu, Tanaka Hozumi and Kimula Kenji, "Paraphrasing Japanese Noun Phrases Using Character-based Indexing", *Proc. of the Second International Workshop on Paraphrasing (IWP2003),* Japan, pp. 80-89. July, 2003.

[5] Yujie Zhang, "Paraphrasing of Chinese Utterances", *Proc. of COLING2002*, Taipei, pp. 1163-1169, August, 2002.

[6] Chengqing Zong, Yujie Zhang, Kazuhide Yamamoto, Masashi Sakamoto and Satoshi Shirai, "Approach to Spoken Chinese Paraphrasing Based on Feature Extraction", *Proc. of NLPRS'01*, Tokyo, Japan, pp. 551-556, November, 2001.

[7] 宗成庆，张玉洁，山本和英，坂本仁，白井谕，"面向口语翻译的汉语语句改写方法"，*汉语语言与计算学报*，12 (1), pp. 63-67, 2002

[8] 张玉洁，山本和英，"汉语语句的自动改写"，*中文信息学报*，Vol. 17(6), pp. 31-38, 2003.

[9] Yujie Zhang and Kazuhide Yamamoto, "Analysis of Chinese Spoken Language for Automatic Paraphrasing", *Proc. of ICCPOL'01*, Seoul, Korea, pp. 290-293, 2001.

[10] Kiyonori Ohtake and Kazuhide Yamamoto, "Applicability Analysis of Corpus-derived Paraphrases toward Example-based Paraphrasing", *Proc. of PACLIC17*, pp. 380-391, 2003.

[11] Xia Fei, The Part-of-Speech Tagging Guideline for the Penn Chinese Treebank (3.0), available at web site http://www.ldc.upenn.edu/ctb, 2000.