Meta Meta

New Light Shed on Chinese Word Segmentation in MT by a Language Investigation

Zhijie Wu

Volume 53, numéro 3, septembre 2008

URI : id.erudit.org/iderudit/019244ar
DOI : 10.7202/019244ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN 0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Zhijie Wu "New Light Shed on Chinese Word Segmentation in MT by a Language Investigation." *Meta* 533 (2008): 630–647. DOI : 10.7202/019244ar

New Light Shed on Chinese Word Segmentation in MT by a Language Investigation

ZHIJIE WU
Nanjing University of Science and Technology, Nanjing, China
eoffrey.zhijie.wu@gmail.com

RÉSUMÉ

La langue chinoise, à la différence des langues occidentales, ne laisse pas d'espace entre deux mots à l'écrit, ce qui pose un problème à la traduction par ordinateur du chinois à l'anglais : comment segmenter les mots en chinois ? Le système de segmentation de mots utilisé actuellement dans la traduction par machine est doté soit d'une orientation linguistique, soit d'une orientation statistique. Cependant, compte tenu du caractère pragmatique de la langue chinoise, les deux genres de système ont des défauts inhérents que l'on n'arrivera pas à effacer. La présente étude propose des solutions pour résoudre le problème de segmentation de mots dans la traduction par machine par une étude langagière composée de deux enquêtes et de huit interviews.

ABSTRACT

The Chinese language, unlike some western languages, is written without a space between any two words, which presents itself as a unique problem in Machine Translation: how to segment words in Chinese? The current word-segmentation systems in Machine Translation are either linguistically-oriented or statistically-oriented. Both types, however, have some innate defects that cannot be overcome due to the pragmatically-oriented feature of the Chinese language. This research aims at addressing the problem of Chinese word segmentation of Machine Translation in light of a language investigation consisting of two surveys and eight interviews.

MOTS-CLÉS/KEYWORDS

Chinese word segmentation, machine translation, language investigation, contextual information, semantic plausibility

1. Introduction

The Chinese language, unlike some western languages, is written without a space between any two words, which presents itself as a unique problem in Machine Translation (hereafter abbreviated as MT): how to segment words and set word boundaries in Chinese? In fact, Chinese word segmentation (hereafter abbreviated to "CWS") is often referred to as the bottleneck for Chinese language understanding and processing. Although its significance has long been recognized, the pioneering research in this field did not begin until the 1980s. Since the first CWS system, i.e., CDWS, was developed in 1983 by the School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, many studies have been conducted and quite a few models have been established (Wang *et al.* 2003). First came the mechanical matching method for CWS, which was employed in the CDWS system. By this method, we first establish a large lexicon that contains (almost) all the possible words in Chinese, and then apply certain rules to divide the input sentence into small chunks,

which are to be compared with the items (i.e., possible words) in the lexicon. If all the chunks match items from the lexicon, then CWS is accomplished. Otherwise, an alternative division of the sentence is carried out and the above process is repeated. The process, however, might be applied several times before an acceptable result can be obtained. Based on this method, there exist some subcategories, such as Maximum Matching and Minimum Matching, and Obverse Matching and Reverse Matching, with the first two divided on the criterion of priority to long words or short words, and the latter pair, on the direction of processing. Among them, the Maximum Reverse Matching Method has been most widely used. This method, however, is notorious for its poor treatment of ambiguity processing of CWS. In order to improve its performance, feature lexicon, binding matrix, and grammar analysis have recently been incorporated into this method. Feature lexicon, by extracting functional words (e.g., "了"), words with affixes (e.g., "老虎"), and words formed by doubling the same character (e.g., "明明白白"), is employed as a kind of pre-processing before applying the mechanical matching process. Binding matrix is used to check results of the mechanical matching process by a grammar matrix and a semantic matrix on the level of phrases. Grammar analysis also applies grammar rules for the purpose of word-segmentation, but it is performed synchronously with the mechanical matching both on the level of phrase and sentence, therefore it is different from the binding grammar matrix. (Liu 2000; Wang *et al.* 2003; Luo *et al.* 1997; Yin 1998.) These, although different from each other, all belong to the category of formal rule-based methods. The subsequent systems based on them do not produce very satisfactory results.

There is another underdeveloped approach for CWS, that is, a statistically-based word-segmentation processing system. It usually employs word frequency and character co-occurrence probability to determine the word boundaries, the only example of which is the system designed by Harbin Industrial University. Although it improves the segmentation for uncommon words, it does not perform well on common words, components of which are very flexible in forming words with other characters, and in most cases multiple in the meaning (Liu 2000; Wang *et al.* 2003). Translation Memory is also in this category, although it seems to be a rather remote relative of the system designed by Harbin Industrial University. Translation Memory deals with CWS (to be more accurate, linguistic chunk segmentation) by following what human translators have already done. At present, it is of very limited use because of the lack of parallel corpora between Chinese and English.

The difference between these two approaches can be described as "deductive" vs. "inductive." "The fundamental difference between them is the source of knowledge that eventually determines the behavior of the system. Deductive […] systems rely on linguists and language engineers, who create or modify sets of rules in accordance with their knowledge, expertise, and intuition" (Carl *et al.* 2000: 223-224) while inductive systems depend on examples, which usually take the form of a corpus. The rules of inductive systems are often derived by the system itself from the examples. Neither of these has so far given a satisfactory response to the increasing need of CWS. The problem is that both approaches, in addition to their obvious advantages, have a number of serious drawbacks and are not well adapted to the peculiarity of the Chinese language.

Why do these systems fail to segment Chinese words effectively? According to the information regarding the designs for these translating systems, most programs

adopt either a linguistically-based or a statistically-based CWS system. However, neither of the segmentation approaches can achieve a very satisfactory result. The rule-based linguistically-oriented CWS systems do not produce very satisfactory results due to the fact that most Chinese words can serve more than one part of speech and have couples of, or even dozens of, meanings, with a single character capable of forming words with many other different characters (and sometimes with itself), coming before or after it. As to a statistically-based CWS system, it cannot solve the problem either. A statistically-based CWS system can only make sure that a certain percentage of word segmentations are all right while leaving the remaining words poorly processed and making ridiculous segmentation mistakes. This approach improves the performance of unusual word segmentation, but does a very poor job concerning common words, components of which are very flexible in forming words with other characters, and in most cases polysemous (Liu 2000; Wang *et al.* 2003).

As some scholars have argued, European languages are mainly syntactically-oriented while Chinese is basically pragmatically-oriented. In an article, Daniel Robertson (2000: 169) refers to Chinese as a "discourse-oriented" language and English as a "syntax-oriented" language. In Yan Huang's book *Anaphora: A Cross-linguistic Study* (2000), he also made the claim very explicitly that Chinese is pragmatically oriented. In other words, pragmatic contexts play an important role in understanding Chinese texts, hence they are significant in CWS. All the above-mentioned systems, however, fail to take pragmatic contextual information into consideration, which largely accounts for their poor word-segmentation performance. Therefore, how to incorporate pragmatic contextual information into CWS systems becomes closely relevant to the problem here. In this research, we plan to investigate how Chinese people carry out CWS in their reading and hope that the knowledge of human's CWS can shed light on this long-standing problem for MT.

2. Surveys, Interviews, Experiments and their Findings

In order to find out how Chinese people segment a Chinese sentence into words and how similarly they carry out this word-segmenting process, we have conducted two survey studies and eight interviews.

2.1 Surveys

2.1.1 The First Survey

2.1.1.1 Subjects

The first survey was carried out on the morning of April 9, 2004. The subjects were 32 third year college students majoring in Business English from an intact class of the Provincial College of Administration in Jiangsu Province. The survey was administered by their teacher during class time. Most subjects responded cooperatively, judging from their eagerness to accomplish the survey questionnaires: when they were given the choice to do the survey either at the beginning or at the end of an instruction period, they cheerfully chose the former option. The basic information about the subjects has been summarized in the following table.

TABLE 1

Respondents' Information in the 1st Survey

	No. of participants	Percent	Valid Percent
male	6	19.4	19.4
female	25	80.6	80.6
Total	31	100.0	100.0

2.1.1.2 Questionnaires

The questionnaire consisted of 24 Chinese sentences for the subjects to perform word-segmentation on. The 24 sentences were adapted from the examples given by *Automatic Word-segmentation and Tagging for Chinese Texts* (Liu 2000: 60-91), which, according to Liu, were some of the sentences drawn from a 5,100,000-Chinese-character corpus, which in turn was built up by randomly selecting news articles from the internet. The sentences chosen in my survey formed twelve pairs. In the two sentences of each pair, they share one identical linguistic chunk, which, however, has different structures and meanings in these two utterances. For example, in the sentence pair "姐妹三人从小学上到中学" and "他从小学戏剧表演," "从小学" is the linguistic chunk they share and refers to "from the primary school" and "(have) learned … since childhood" respectively and should be segmented differently, i.e., "从‖小学" in the first case while "从小‖学" in the latter. However, the two sentences in each pair were not listed together; rather, they were dispersed randomly. The subjects were asked to complete the task of word-segmentation for all these sentences following the instruction and illustration in the questionnaire.

2.1.1.3 Data Collection

Three Chinese professors from Nanjing University and National University of Singapore were asked to set a standard segmentation for these sentences.[1] Considering that semantically or structurally ambiguous linguistic chunks present the greatest difficulty for a machine to perform word-segmentation, and also for the sake of convenience, we only graded the identical part in each sentence pair. A 3-point scoring scale was used to check the subjects' performance against the standard segmentation, with "3" for the standard answer, "1" for the wrong answer and "2" referring to the kind of answer that could not be graded. The reason that some answers are not gradable is due to the fact that some subjects segmented the sentence into some linguistic chunks larger than words (usually a word group), which need to be further divided in order to see whether his interpretation was right or wrong. This problem will be addressed in the follow-up survey.

2.1.1.4 Data Analysis

The statistical program SPSS (Version 12.0) was used to analyze the data obtained from the first survey. The following is the descriptive statistics of the survey.

Table 2

Descriptive Statistics of the 1st Survey

	N	Mean
score of segmentation for sentence 17	31	1.4516
score of segmentation for sentence 24	31	1.8871
score of segmentation for sentence 11	31	2.0323
score of segmentation for sentence 3	31	2.1290
score of segmentation for sentence 23	31	2.2903
score of segmentation for sentence 18	31	2.5806
score of segmentation for sentence 21	31	2.5806
score of segmentation for sentence 14	31	2.6129
score of segmentation for sentence 22	31	2.6774
score of segmentation for sentence 7	31	2.8065
score of segmentation for sentence 19	31	2.8065
score of segmentation for sentence 20	31	2.8065
score of segmentation for sentence 2	31	2.8387
score of segmentation for sentence 9	31	2.8710
score of segmentation for sentence 5	31	2.9032
score of segmentation for sentence 13	31	2.9032
score of segmentation for sentence 15	31	2.9355
score of segmentation for sentence 16	31	2.9355
score of segmentation for sentence 1	31	2.9355
score of segmentation for sentence 12	31	2.9355
score of segmentation for sentence 10	31	2.9355
score of segmentation for sentence 4	31	2.9677
score of segmentation for sentence 6	31	2.9677
score of segmentation for sentence 8	31	3.0000
Valid N (listwise)	31	

As we can see from the above list, the results of the survey study, instead of being homogenous, turned out to be rather complicated, with the lowest score being 1.4516 and the highest one 3.000. The perplexing results could be attributed to several intervening factors, of which the most obvious and influential one was the subjects' different perspectives towards the definition of **word**. An extreme example would be a couple of students counting "固定资产更新改造" and/or "农民个人利益" as a word. This viewpoint is further confirmed by the interviews discussed later. There were still some other factors to complicate the matter which were directly related to the wording of questionnaire items. First, one of the sentences (其实质不好，量再多也没用) is ambiguous in itself and the common part "其实质" can be segmented in two ways (either 其实‖质 or 其‖实质).[2] Both types of segmentation are plausible. In fact, most respondents had chosen the less desirable segmentation (其‖实质) rather than the standard solution (其实‖质), thus the scores for segmenting this sentence is the lowest and also the only one below 1.5. Sentences 24 and 11 were the next two lowest

scored items. After careful investigation, we found that they contained "一起" and "十分" respectively, which most subjects did not further divide into the standard answers "一" and "起," and "十" and "分." The follow-up survey and the subsequent interviews showed that they were well aware that the two "一起" were quite different from each other in the sentence pair "这是一起领导干部违纪事件" and "少年儿童一起拉小提琴." This was also the case of "十分" in the sentence pair "现在差十分七点" and "校园环境十分优美." Among the remaining 21 sentences, the scores for 19 sentences were above 2.5. So if some intervening factors had been excluded, the results actually would be very homogenous.

2.1.2 The Follow-up Survey

2.1.2.1 Subjects

In order to pin down the moot score "2" (i.e., the answers that could not be graded due to the fact that some subjects segmented the relevant sentence into some linguistic chunks larger than words) and check the subjects' understanding of the sentences under discussion, a follow-up split survey study was conducted. For some reason it was administered on two different days, i.e., on April 16 and April 23, 2004 respectively. The subjects were the same as those in the first survey but the situation was a bit different. Among the 31 subjects, 28 subjects were present on the morning of April 16 and accepted the follow-up survey. The remaining three were absent on April 16 and didn't complete the survey until the morning of April 23. In both situations, the importance of the study was highlighted by their teacher and the questionnaires were completed in class. The basic information about the subjects has been summarized in the following table.

Table 3

Respondents' Information in the Follow-up Survey

	No. of participants	Percent	Valid Percent
male	6	19.4	19.4
female	25	80.6	80.6
Total	31	100.0	100.0

2.1.2.2 Questionnaires

The questionnaire was made up of the same 24 Chinese sentences as those in the first survey study, which were in turn grouped into 12 sentence pairs. The two sentences in each pair had one identical linguistic chunk and the subjects were asked to judge whether the linguistic chunk was of the same structure and meaning or of different structures and meanings in these two sentences. The subjects were required to accomplish two tasks. First, they were to choose an answer from the five options varying from **Identical** to **Very Similar** to **Undecided** to **Very Different** to **Totally Different** according to their judgment (Task I). After that, they were asked to state briefly why they chose such an answer (Task II). The items of questions for each subject to answer were different from each other. The questions were specified according to how they performed in the previous survey. A). The subjects were to accomplish those items that were related to the sentences of the first survey for which their segmentation had

got the score "2." They only needed to choose an answer for their questions (Task I) and this part was compulsory. B). They were welcome to finish the items that contained the sentences for which their segmentation in the first survey had got the score "1." In this part, they first conducted multi-choice for these items, after which they were asked to state the reason(s) for their choice briefly (Task I and II). They might choose to or not to finish this part. It was optional.

2.1.2.3 Data Collection

As I have pointed out before, the follow-up survey was mainly applied to addressing the moot score "2" (the compulsory part), which was the kind of answer that could not be graded because the subject segmented the sentence into some linguistic chunks larger than words (usually a word group). Consequently, the results of the second survey were not recorded separately. Instead, they were incorporated into the first survey with the original results revised. The revising rule is as follows. If the subjects choose the answer "**Identical"** or "**Very Similar**," their original score "2" will be changed into "1." If they select the answer "**Undecided**," the moot score "2" will remain the same. And if they opt for "**Very Different**" or "**Totally Different**," "2" will be turned into "3."

As for the other part of the survey (the optional part), some of the results were also used to revise the statistics obtained from the first survey. The revising principle is like this. 1). If their choice is consistent with what they did in the first survey, no change will be made. 2). If they select an answer contrary to the one they chose in the first survey (usually other than "Undecided") and their interpretation is correct, their score for the corresponding item will be changed from "1" to "3." 3). If they choose "Undecided," the score will be changed into "2." There are still some other results (mainly the interpretative part of their answers) which are to be employed in the later qualitative analysis.

It should also be noted that there is one assumption behind the revising work. That is, the subjects should have no problem in interpreting the meaning of sentences since all of them are native speakers receiving their higher education. Their understanding might vary, but not to such a significant extent that it should be taken into account. And this assumption was somewhat supported by the subsequent interviews.

2.1.2.4 Data Analysis

In the compulsory part of the survey, there are altogether 96 times of questions. (Different subjects may have the same question, which will be counted as several times of questions rather than one.) All of these, as expected, were answered by the subjects. 83 out of these 96 are consistent with the standard answers. The right percentage is as high as 86.5%. Here, the truly ambiguous sentence "其实质不好，量再多也没用" appeared twice. If we discard these two, the correct percentage increases to more than 87.2% (82/94).

For the optional part, we have altogether 44 times of questions, 40 of which were answered. The respondent rate is about 90.9%. Among these 40, only 22 conform to the reference answers. But if we look at the results more carefully, we find that 22 (bearing no relationship to the above number 22) out of 44 relate to the truly ambiguous sentence "其实质不好，量再多也没用." If these 22 are discarded, the right percentage of answered items will be more than 84.2% (16/19).

After we incorporated the above results into those of the first survey, we get the following descriptive statistics.

Table 3

Descriptive Statistics of the Two Surveys

	N	Mean
score of segmentation for sentence 1	31	3.0000
score of segmentation for sentence 2	31	3.0000
score of segmentation for sentence 3	31	2.9355
score of segmentation for sentence 4	31	3.0000
score of segmentation for sentence 5	31	3.0000
score of segmentation for sentence 6	31	3.0000
score of segmentation for sentence 7	31	2.9355
score of segmentation for sentence 8	31	3.0000
score of segmentation for sentence 9	31	3.0000
score of segmentation for sentence 10	31	3.0000
score of segmentation for sentence 11	31	3.0000
score of segmentation for sentence 12	31	3.0000
score of segmentation for sentence 13	31	3.0000
score of segmentation for sentence 14	31	2.6129
score of segmentation for sentence 15	31	2.9355
score of segmentation for sentence 16	31	3.0000
score of segmentation for sentence 17	31	1.9032
score of segmentation for sentence 18	31	2.7097
score of segmentation for sentence 19	31	2.9355
score of segmentation for sentence 20	31	3.0000
score of segmentation for sentence 21	31	3.0000
score of segmentation for sentence 22	31	3.0000
score of segmentation for sentence 23	31	2.6129
score of segmentation for sentence 24	31	3.0000
Valid N (listwise)	31	

This is really an impressive table. Among all the 24 sentences, 16 get the full score "3" and altogether 20 are above the score 2.9 with only one – the truly ambiguous sentence 17 – under the score of 2.5. And from the results of these two surveys, we may safely claim that different people carry out the CWS process with great similarity.

The striking change from Table 2 to Table 4 also shows that real time pressure exerts some influence on their word-segmentation performance. Better results were obtained when the subjects were given the opportunity to check and revise their segmentation.

In the process of these two surveys, there was another phenomenon worth mentioning (although some people may think it is commonplace): usually the subjects did not write down their answers until they finished reading the whole sentence.

Even if they did write down their answers, they might revise their segmentation of the first part of sentence when they obtained new information from the second part that was contradictory to their former understanding. Nine subjects just crossed out their former answers (instead of erasing them) and therefore left traces on their questionnaires. Further examination of these surveys revealed that the revised sentences were usually "与其他点灯，不如我放火" and "其实质不好，量再多也没用." It is really no coincidence that these two were the revised sentences, because these two need the information in the second part (不如 and 量) to decide the CWS in the first part (与其‖他 and 其实‖质). This phenomenon shows that human beings' word-segmentation is a bi-directional process and cannot be settled at one go, which is further confirmed by the findings in the interviews.

2.2 Interviews

Following the two surveys were eight interviews. They were conducted on the basis of the surveys and aimed to find out how the subjects segmented the sentences in the surveys.

2.2.1 Subjects

Eight subjects were randomly selected from the same class and they were seven females and one male. The general information about the subjects and their interviews has been summarized as follows.

Table 4

General Information about Interviews

Interviewees	Gender	Length of the interview	Date
A	Female	00:04:20	04/09
B	Male	00:07:18	04/09
C	Female	00:08:50	04/09
D	Female	00:02:35	04/16
E	Female	00:02:05	04/16
F	Female	00:02:48	04/16
G	Female	00:03:10	04/16
H	Female	00:06:42	04/16

2.2.2 Measures

The interviews were based on the two surveys and covered all the 12 pairs of sentences. The subjects were required to recall their decision-making process concerning whether the same linguistic chunk in a certain sentence pair was of different structures and meanings (or of the same structure and meaning, if their answers were so decided). Usually, the interviewees were asked to explain why they segmented the sentences in such a way. Each interviewee had to explain his/her CWS for 2 to 4 sentence pairs. Altogether, 108 times of sentences (54 pairs of sentences) were discussed (One sentence may be discussed several times and counted as such.).

2.2.3 Data Collection

All the interviews were recorded and transcribed. Upon careful examination of all these interviews, the interviewees' answers were grouped into two categories: they are either contextual/semantic or structural/syntactical. For example, an answer like "(My decision was based on) the whole sentence. … '与其他点灯' echoes with '不如我放火.' Well, it's '与其' followed by '不如.' '与其' what what, '不如' what what. (整个句子啊。……与其他点灯，不如我放火，相对应嘛。然后，它就是'与其'，然后'不如'。与其怎么样，不如怎么样。)" will be counted as a decision made out of the contextual information, whereas "'他将来太原工作,' I think, just means '他会来.' (他将来太原工作，我考虑到，就是说，是他会来。)" will be regarded as a decision relying on the semantic clues. The structural cues and the syntactical hints will be considered as the main source for word-segmentation in the following two cases respectively: "In '与其它领导,' '与' means '什么和什么.' '与' (…) just means 'A和B' ('与其它领导', '与'是'什么和什么'。'与'……就是'A和B'嘛。)" and "Its subject is '变价收入.' '应,' is served as an adverb, isn't it? Then, '用于' is a verb, used as the predicate. (它主语是'变价收入'。'应'，是作那个状语吧。然后'用于'是谓语动词。)." The contextual clues and the semantic information will be considered together, as they are often related and overlapped, such as those in "Then, here comes '七点.' Then it might be time, as it occurred to me that this is '十分.' It is a concept about time. (然后它讲'七点'嘛，然后它可能就是，我又想它是'十分'，就是一个时间嘛。它是一个时间上的概念。). The same treatment goes for the structural and the syntactical information.

2.2.4 Data Analysis

In all the 54 cases of word-segmentation explanation, most interviewees reported that they made their decisions upon contextual and semantic information of the sentences, which made up about 83.3% of all the cases. Only in 9 cases did they report the structurally and/or syntactically based word-segmentation.

In the data analysis, I also differentiate the main criteria from the minor ones since quite a few interviewees made use of several kinds of information in their CWS of one sentence. These minor ones often played a supportive role in their word-segmentation. In 27 cases, the interviewees used these supportive clues. Among them, 14 supportive clues were contextual and/or semantic information while 13 belonged to the structural and/or syntactical category.

The results strongly indicate that people usually carry out the word-segmentation process against contextual and semantic information available in the text rather than the structural information of the text. In other words, the most frequently used word-segmenting strategy by human is to find contextual/semantic information. By contrast, the structural/syntactical information is employed mainly to support their CWS. It only plays a supportive role in their CWS.

Another phenomenon figures prominently in all these interviews: the interviewees gave their interpretation and explanation of the sentences at all times, which implies that their criterion for word-segmentation is semantic plausibility.

Although structural/syntactical information played a limited role in the word-segmentation decision-making process, we could still see an explicit inclination of the interviewees to resort to this kind of information. This tendency, we believe, has something to do with their major. As we have pointed out, the subjects were third

year college students majoring in Business English from an intact class of the Provincial College of Administration in Jiangsu, China. They have received education in English for quite a few years (at least nine years, if their study of English in high school was included). And as the grammar-oriented teaching method prevailed in most high schools of China in the past decade, it is no wonder that they made great efforts in employing this kind of grammar knowledge in their CWS as a result of language transfer, though unsuccessfully in most cases. One particular interviewee even made explicit reference to her English knowledge: "I feel that it is like English. … I feel if it were English, an adverb would do in this context. (我感觉，就像这个英语。……感觉这个英语，其中是副词就行。)" In fact, she stated again and again that she carried out the CWS according to parts of speech of the words in the sentence. "My judgment is that they play different roles. What are their parts of speech? Different (parts of speech). It was in this way that I made my judgment. (我这个判断就是说，它(们)成分不一样。……是怎样一个词性呢？不同。我就这样判断的。)." "I always (made my decision) by its role in sentences. Just the role, the role it played in a sentence, (served as the source) for my word segmentation. (我都是根据它的成分。就是那个成分，句子里作的成分，来切分的。)" Quite ironically, she was actually not very successful in her attempts to do so. When she was asked why these same Chinese characters are of different parts of speech, she was unable to give any answers: "I don't know. (我也不知道)." Then, she tried to resort to her feel of language: "Anyway, it's the feel. (反正就是感觉。)" Later, she unconsciously made use of contextual and semantic approaches in her CWS. "(We cannot link it with this '上.' ('它(指'大红马')跟这个'上'联系不起来。)" "He will go to that place and work there. ('它将会到那个地方去工作。)" "'将来会更完善' is, just means '以后会更好,' isn't it? ('将来会更完善'，就是表示的就是'以后会更好'吧。)") If we carefully examine what she said, we shall find out that in most cases she actually made use of the strategies of both contextual clues and semantic plausibility rather than "parts of speech" to carry out the CWS. Part of speech was only applied to check the semantic plausibility. And it is not surprising that she eventually resorted to contextual and semantic approaches, because a Chinese character usually has no inherent and unique "part of speech." In the 9 cases of the structurally and/or syntactically based word-segmentation, she claimed five of them. If these "pseudo-cases" of structurally oriented word-segmentation were ruled out, we can safely claim that in the overwhelming majority of cases the interviewees carried out their word-segmentation process from a contextual and semantic perspective.

Another finding from these interviews is that the contextual information employed in segmenting a certain word may not be the immediate context of the prospective word (i.e., a linguistic chunk that is expected to be a word, but needs further confirmation), which may go either before or after it. Immediate context here refers to one or several characters that go immediately before or after the prospective word. Take the sentence "与其他点灯，不如我放火" as an example. For the linguistic chunk "与其他," "点" or "点灯" is its immediate context, while "不如" will not be considered as an immediate context of it. Besides, the contextual clue may appear either before or after the prospective word. The prospective word and its clue often form a semantic pair and echo each other. This finding confirmed our hypothesis stated in the previous part that people's word-segmentation is a bi-directional, dynamic process rather than a one-directional treatment. It cannot be settled at one go.

Additionally, we came across the problem of the definition of WORD in several interviews, which strengthened the finding in the first survey. In one case, the interviewee said "I think '农民个人利益' should be counted as a word. (我觉得'农民个人利益'应该算是一个词。)" She did not differenciate between a word and a word group. In another case, although the interviewee differentiated the two "一起" in the sentence pair "这是一起领导干部违纪事件" and "少年儿童一起拉小提琴," he did not further divide "一起" of sentence 24 into the standard segmentation "一" and "起." These interviews show that people tend to have different opinions about what a word is, and confirm some previous findings. (See Wang Li 2003)

2.3 Experiments

In order to see how and how well MT programs carry out the word-segmentation process, we conducted these experiments. We intended to compare their segmentation with that of human beings, hoping to draw inspiration from the comparison so as to improve the current CWS programs.

2.3.1 Materials

The materials used in the experiments remain the same as those employed in the surveys and interviews: they were twelve pairs of sentences adapted from the examples given in *Automatic Word-segmentation and Tagging for Chinese Texts* (See Liu 2000: 60-91). The purpose of using the same textual materials is to make experimental results and those from surveys and interviews more comparable.

2.3.2 Instruments

Two famous MT programs from Mainland China and an automatic word-segmentation program from Taiwan China were employed here to process the above-mentioned sentences. It should be pointed out that we had intended to use more MT programs in the experiments. However, most MT programs only offer English-to-Chinese (E-C) translation service, such as IBM Translator (IBM翻译家), FastAIT (东方快车), Dr. Eye (译典通), which may be due to the reason that Chinese-to-English (C-E) MT techniques are still underdeveloped and lag far behind E-C MT techniques (CWS is a big contributing factor to this state of underdevelopment.). In order to compensate for it, we not only made use of the two C-E MT programs available but also conducted our experiment on an automatic word-segmentation program. Fortunately, these three programs were from three famous companies or agencies and were quite representative: one was a product of the Kingsoft, one was from the Hero Corporation, and the third one was developed by Chinese Knowledge Information Processing Group (CKIP), Academia Sinica, Taiwan, China.[3]

2.3.3 Data Collection

Although the experiments were carried out on the level of sentence, our attention still focused on those linguistic chunks shared by the two sentences in each pair. The data collected were of two types. One was about whether the programs rightly segmented this linguistic chunk into words, and the other, about whether the programs correctly translated this part. (The automatic word-segmentation program could only supply the first kind of data since it does not offer the translation service.)

2.3.4 Data Analysis

The results show that the right segmentation rate is not very high. In fact, the experiment outputs indicate that these programs did a somewhat poor segmentation job, which well justified the significance and necessity of this research. Of all the 24 linguistic chunks, the Kingsoft program has correctly segmented 13, the Hero program, 16 and the automatic word-segmentation program, 2. If we just look at these scores, we might say that the performance of the first two is not too bad, since they have worked out more than a half. But actually this is not the case. If we examine these results in the unit of sentence pair, we might get a totally different view – and a more accurate one, we think. The Kingsoft program was "consistent" in its treatment with these linguistic chunks. It did not change its segmentation in 11 out of 12 sentence pairs, lacking the ability to take different contexts into consideration. The only exception is "个人" in the sentence pair "谁也不能损害农民个人利益" and "他一个人睡在屋里." The Hero program has improved slightly, but only slightly. It was only able to vary its segmentation according to the different linguistic environments in 4 sentence pairs while leaving the other 8 pairs wrongly segmented. If the sentence pair containing the truly ambiguous sentence "其实质不好，量再多也没用" is discarded, their ability to take adaptive segmentation for the same linguistic chunk in different environments is rather low, achieving a correct segmentation rate of 9.1% and 36.3% respectively. The performance of the automatic word-segmentation program, as shown in the above, is the worst. It only correctly segmented two of the twenty-four linguistic chunks under study. If we look at its segmentation in the unit of sentence pair, we can see that it has almost never varied its segmentation although the context for the linguistic chunk has changed. The difference of their performance may be partly attributed to the different language conventions practised in Mainland China and Taiwan China. But the fact is that the language difference is not so great as to cause such a significant disparity of word-segmentation. The reason for such a poor performance, we assume, still rests largely on the underdeveloped word-segmentation techniques.

The translation of the parts under study in each sentence pair was also carefully examined. The Kingsoft MT program wrongly translated 14 of those 24 linguistic chunks whereas the Hero MT program gave incorrect rendering to 12 of them. Wrong segmentation of each part necessarily resulted in the wrong translation of it, which is easily understandable. The translation mistakes caused by wrong CWS make up 78.6% (11/14) and 66.7% (8/12) of all the translation mistakes for the Kingsoft program and the Hero program respectively. Besides, we also find some linguistic chunks which were correctly segmented but incorrectly translated. The problem lies in wrong selection of the constituent words' meanings. As we know, a Chinese word may have several meanings. Which one is appropriate depends on the context in which the word appears. Therefore, the contextual information is also important and necessary in the selection of one proper meaning from many possible meanings.

It may be worth mentioning that apart from the linguistic chunks under examination, there are usually some other parts wrongly segmented and translated. Altogether, these two MT programs have only a couple of acceptable translated sentences, which are just some very simple sentences, such as "The campus environment is very beautiful" and "It is 10 minutes to 7 o'clock now."

3. Findings and Their Implications

In this part, we will discuss the results obtained from our surveys and interviews. This discussion will be carried out in light of the results of machine translation experiments and try to foreground those findings that show the methodological and strategic difference of word-segmentation between humans and machines, in the hope that the knowledge of humans' unique word-segmentation process may shed light on CWS techniques and contribute to CWS model-designing. In other words, this research aims to draw inspiration from the cognitive process of human's CWS and teach machines how to do word-segmentation.

3.1 Potentiality for Machines to Carry Out Word Segmentation

From the surveys and interviews, we find out that people have achieved a relatively homogeneous word-segmentation result, obtaining an almost identical understanding. This shows that there is a standard way of segmenting and interpreting a certain sentence. However, we have to admit that there are some ambiguous sentences which may be open to different interpretations. But even for these ambiguous sentences, we can arrive at a consensus of the possible and plausible interpretations of them. It is not the case that we could interpret these sentences at will. The existence of **right** interpretation and segmentation is very important, which implies that it has potentiality for machines to carry out word-segmentation. Otherwise, if people could not agree with each other on the correct word-segmentation and interpretation, how can we expect machines to carry out this process and what is the criterion for their processing?

3.2 Contextual Information in Word Segmentation

As indicated by the surveys and interviews, the most frequently used word-segmentation strategy by humans is to find contextual information. This kind of information is the main source for a person to carry out CWS.

More than that, the contextual information a person employs in his/her CWS is not limited to the immediate context. In other words, the contextual clues for human CWS may be several characters or words apart from the prospective word.

By contrast, most of the current machine translation systems available on the market can only make use of structural clues. According to the information with regard to the designs for these translating systems, they almost make no use of contextual information. There are a few exceptional systems that do use contextual information in their CWS. The CWS engines employed in these exceptional translation systems are usually statistically-based CWS systems, which employ word frequency and character co-occurrence probability to determine word boundaries. We should admit that word frequency and character co-occurrence are a kind of contextual information. However, this kind of contextual information is very different from the contextual clues that human beings makes use of in their CWS: it is only immediate context and cannot be separated from the prospective word by other characters, words, phrases, clauses or sentences.

Furthermore, the contextual information humans employ in CWS does not necessarily occur before or after the prospective word. It may take either position. This point has been supported by both the surveys and interviews. In the surveys,

we observed that some subjects revised their segmentation of the first part of sentence if they obtained new information from the second part that was contradictory to their former understanding. Their CWS underwent several stages, such as reading, tentative segmenting, expectation, more reading, more tentative segmenting, confirming or revising the previous tentative segmentation, more expectation…. The interviewees usually segmented the sentence from the left to the right, but they also looked back from time to time, checking and revising. This is really a dynamic process and cannot be settled at one go. The interviews also produced hard evidence for the view of human's two-way processing. For example, "与其" and "不如" appeared in different clauses of a complex sentence, but they echo with and provide segmentation clues for each other. They formed a kind of semantic link, with the former being the reason to pick the latter up as a word, and the latter being the reason to single out the former as a word. This is also the case of "质" and "量"in the sentence "其实质不好，量再多也没用."

This, again, contrasts sharply with what is going on in MT programs. Most MT programs only apply a one-way segmenting process, either Obverse Matching or Reverse Matching, as mentioned in the introduction. Compared with human's dynamic bi-directional CWS process, it is no wonder that this kind of unidirectional mechanical CWS process fails to do a good word-segmentation job.

3.3 Semantic Plausibility as the Criterion of Word Segmentation

The interviewees, as we have already discussed in the Interviews section, made considerable use of interpretation. This kind of semantic judgment is abundant in all interviews. This rather clearly informs us of their criterion for CWS, that is, semantic plausibility: if we have segmented the text in a way that the results are semantically reasonable and plausible, then "Bingo!" The segmentation is all right. Otherwise, we have to try again.

CWS systems in Machine Translation, different from the cognition of a person, can only deal with the lexical possibility, i.e., whether those linguistic chunks are possible words. This often leads to word-segmentation mistakes, and as a result accounts for some of the most ridiculous translation errors of MT programs. However, there is little hope of dramatic change of this situation because semantic plausibility entails inference and interpretation, an ability that seems easy to a human being but is extremely difficult for a machine. We know interpretation is a kind of creation in nature. A computer, however, is "fundamentally just a device for following rules, mechanically and literally, albeit with considerable speed and precision. Rule following can produce a kind of creativity, but not the kind of creativity required for interpretation." (Somers 2003: 120) However, does this mean computers can do absolutely nothing about it? No. There is still something a computer could do – inference, a necessary and integral part of interpretation. Machines are capable of some elementary inference. For example, if computers find that linguistic chunks "与其" and "不如" turn up in a sentence at the same time, they could pick them up first and count them as words. (With time and development, we think, we human beings will make it possible for a computer to interpret an utterance or text.)

The function of semantic plausibility, in fact, is more than an aim of CWS. Apart from the interviewees' judgment of word-segmentation for the whole sentence, the

semantic information had also been applied much earlier in their segmentation of sentential fragments, in their tentative CWS of part of the sentence under discussion. This is to say, semantic consideration is not only an end but also a means in the process of CWS. The semantic information is also an important clue for human CWS. Perceived from this perspective, the CWS process becomes a process of trials and errors: we try a possible segmentation and interpretation of part of a sentence, then a possible segmentation and interpretation of the second part of the sentence (of course, there might be more parts if the sentence is long), then gather up the partial interpretations to get a holistic interpretation of the sentence: we will select this segmentation and interpretation if it seems semantically plausible; otherwise, we will reject and revise this segmentation and interpretation. From the above depiction of word-segmentation process, we may see that word-segmentation and interpretation go hand in hand. They are closely interrelated with each other.

It is no coincidence that semantics acts as both a means and an end of CWS. First, this is because meaning, or more accurately, semantic plausibility, is the ultimate goal of natural language understanding, and consequently the ultimate goal of CWS. Second, semantic plausibility is also the criterion of understanding any part of natural language, therefore it accompanies the whole process of natural language understanding and appears both as the "starting point" and the "finish line" of CWS.

Meaning is the main theme throughout the whole CWS and natural language understanding process. As far as translation is concerned, it plays an even more important role. Meaning is not only the goal of source text interpretation, but also the content that is transferred into the target language. Just as Nida puts it, "Translation means translating meaning (Nida *et al.* 1986)."

3.4 Word Segmentation: a Means Rather Than a Purpose

From the above discussion, it should be obvious at this point that CWS itself is not the purpose. It is only part of the natural language understanding process.

Some scholars argue that we should standardize the CWS processing so that it could be applied to information retrieval, natural language processing, MT and other related areas. We should say this is a good proposition, and has its obvious advantages. First, it would be very economical and save much time and energy if one single CWS system could cater to the needs of all these different areas. Secondly, it would be much easier to repair or adjust the CWS system since it is independent of other parts. The repair, adjustment or even substitution of CWS subsystem would have minimum impact on the other parts of the whole macro system.

We, however, hold a different opinion although we are well aware of the advantages of such a standardized word-segmentation system. (Our concern here mainly goes to the CWS subsystem in MT programs.) We think that CWS cannot be separated from other parts since it is an integral part of natural language understanding. CWS is not a self-sufficient system and requires relevant knowledge and information to perform its task. The process of CWS is at the same time a process of natural language understanding and processing. The result of this process, i.e., the segmented text, does not carry all the information. For example, if in the sentence "其实质不好，量再多也没用," we make use of the semantic link "实" and "质" to segment each other, meanings of these two words can be pinned down in this context, too. However,

if only the word-segmentation result is used, we still need to tell computers to select a proper meaning for each of them, which is likely to be a wrong choice since each of them has several meanings. Therefore, if only the segmented text is used, much useful information will be lost. This is really a great waste, and sometimes this kind of waste is unrecoverable. We propose to retain all the information until it is used in the later translating part, such as meaning selection.

At the same time, as we have pointed out before, we are fully aware of the merits of a standardized word-segmentation system. In fact, we could still have some standardized "spare parts" so that the merits could be retained. But the standardized part is not the whole word-segmentation system. Rather, we could standardize some smaller parts, such as the lexicon, and the knowledge-based word-segmentation rules. Consequently the question becomes what is the optimal size of "standardized spare parts" and which are the most desirable parts to get standardized, rather than whether we should standardize or not.

3.1.5 Implications

In the previous part, we discussed and compared the results from surveys, interviews and experiments. The major findings can be summarized as such: human beings achieve a relatively homogeneous word-segmentation result, obtaining almost identical understanding. Their most frequently used word-segmentation strategy is to find semantic and/or contextual information, which is not restricted to immediate context and can appear before or after the prospective word. And their criterion for CWS is semantic plausibility. The current CWS systems in MT, by contrast, seldom employ contextual information. Instead, they usually make use of structural clues and in most cases leave semantics unconsidered, which largely accounts for their poor word-segmentation performance. Additionally, CWS systems in MT are usually independent of other parts, wasting part of the information obtained from the CWS process.

Based on these findings, some implications are drawn in the hope that they will be useful in future MT system designing, especially regarding the part of CWS.

First and foremost, if pragmatic and contextual information is incorporated into word-segmentation systems, the CWS problem will be more satisfactorily resolved and machine translation performance will be greatly improved. Besides, the contextual information employed in CWS should not be restricted to immediate context. As a result, the processing unit should be enlarged so that more contextual information can be taken into consideration.

Second, semantics should play a more important part in the CWS system. Many semantically-oriented rules should be incorporated. The structurally-oriented rules might be retained but they must be relegated to the marginal status.

Third, since word-segmentation is a dynamic process and cannot be settled at one go, the one-way mechanical processing approach should be discarded in favor of a bi-directional one.

Last but not least, some of the processing information during CWS contains useful information which should be retained for later use, especially in the part of meaning selection. Therefore, it is advantageous of us not to make CWS as a standardized part. Instead, we could standardize some smaller parts (such as the lexicon and the knowledge-based word-segmentation rules), allowing the CWS system to interact with other parts of MT.

NOTES

1. Professor Xu Daming (徐大明) from Nanjing University has kindly helped me to establish the standard segmentation. But he also emphasized that "linguists' judgment are [sic] in fact inferior to laymen's in terms of 'intuition.' What linguists can do is to be detached from 'being a native speaker' and find objective ways in observing people speaking." —Quoted from his email sent to me on April 14, 2004.
2. Some subjects were well aware that the sentence "其实质不好，量再多也没用" is ambiguous. In the subsequent follow-up survey, one subject stated openly that "(其实质)既能指它的实质，也能指其实‖质."
3. Hereby I express my gratitude to Kingsoft Corporation, Hero Corporation and CKIP. (谨此向金山快译、豪杰译霸和中央研究院词库小组致谢。)

REFERENCES

Carl, M. *et al.* (2000): "Towards a Dynamic Linkage of Example-based and Rule-based Machine Translation," *Machine Translation* 15, pp. 223-257.

Huang, Y. (2000): *Anaphora: A Cross-linguistic Study* (Oxford Studies in Typology and Linguistic Theory), Oxford, Oxford University Press.

Liu, K. (2000): *Automatic Word-segmentation and Tagging for Chinese Texts* (In Simplified Chinese), Beijing, The Commercial Press.

Luo, Z. *et al.* (1997): "A Review of the Study of Chinese Automatic Segmentation" (In Simplified Chinese), *Journal of Zhejiang University*, 31-3, pp. 306-312. (In Simplified Chinese)

Nida, E. A. and J. de Waard (1986): *From One Language to Another*, Nashville: Thomas Nelson.

Robertson, D. (2000): "Variability in the Use of the English Article System by Chinese Learners of English," *Second Language Research* 16-2, pp. 135-172.

Somers, H. (ed.) (2003): *Computers and Translation: A Translator's Guide*, Philadelphia, John Benjamins Publishing Company.

Wang, K. *et al.* (2003): "The Main Techniques in Chinese Word-Segmentation and Its Prospect of Application" (In Simplified Chinese), *Communications Technology* 138, pp. 12-15.

Wang, L. (2003): *A Sociolinguistic Study of Chinese Words* (In Simplified Chinese), Beijing, The Commercial Press.

Yin, J. (1998): "Automatic Word Segmentation Methods for Chinese Language" (In Simplified Chinese), *Computer Engineering and Science* 20-3, pp. 60-66.