Meta
Journal des traducteurs
Translators' Journal

META

A Cognitive Model of Chinese Word Segmentation for Machine Translation

Zhijie Wu

Volume 56, numéro 3, septembre 2011

URI : https://id.erudit.org/iderudit/1008337ar
DOI : https://doi.org/10.7202/1008337ar

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

Wu, Z. (2011). A Cognitive Model of Chinese Word Segmentation for Machine Translation. *Meta*, *56*(3), 631–644. https://doi.org/10.7202/1008337ar

Résumé de l'article

À la différence de l'anglais, la langue chinoise ne marque pas la délimitation entre les mots. C'est pourquoi la segmentation du chinois constitue l'obstacle principal de la traduction automatique vers l'anglais. Actuellement, les méthodes de segmentation en traduction automatique sont soumises à des règles linguistiques ou font appel à des analyses statistiques. Le chinois, toutefois, présente des caractéristiques pragmatiques très fortes, ce qui explique l'échec des stratégies actuelles. Nous avons réalisé une étude constituée de deux enquêtes et de huit entrevues visant à déterminer comment les Chinois segmentent une phrase dans leur langue en situation de lecture. Sur la base des résultats obtenus, nous avons mis au point un nouveau modèle de segmentation lexicale visant à résoudre la question de la segmentation en traduction automatique sous un angle cognitif.

A Cognitive Model of Chinese Word Segmentation for Machine Translation*

ZHIJIE WU
Nanjing University of Science and Technology, Nanjing, China
University of California, Berkeley, USA
geoffrey.zhijie.wu@gmail.com

RÉSUMÉ

À la différence de l'anglais, la langue chinoise ne marque pas la délimitation entre les mots. C'est pourquoi la segmentation du chinois constitue l'obstacle principal de la traduction automatique vers l'anglais. Actuellement, les méthodes de segmentation en traduction automatique sont soumises à des règles linguistiques ou font appel à des analyses statistiques. Le chinois, toutefois, présente des caractéristiques pragmatiques très fortes, ce qui explique l'échec des stratégies actuelles. Nous avons réalisé une étude constituée de deux enquêtes et de huit entrevues visant à déterminer comment les Chinois segmentent une phrase dans leur langue en situation de lecture. Sur la base des résultats obtenus, nous avons mis au point un nouveau modèle de segmentation lexicale visant à résoudre la question de la segmentation en traduction automatique sous un angle cognitif.

ABSTRACT

The Chinese language, unlike English, is written without marked word boundaries, and Chinese word segmentation is often referred to as the bottleneck for Chinese-English machine translation. The current word-segmentation systems in machine translation are either linguistically-oriented or statistically-oriented. Chinese, however, is a pragmatically-oriented language, which explains why the existing Chinese word segmentation systems in machine translation are not successful in dealing with the language. Based on a language investigation consisting of two surveys and eight interviews, and its findings concerning how Chinese people segment a Chinese sentence into words in their reading, we have developed a new word-segmentation model, aiming to address the word-segmentation problem in machine translation from a cognitive perspective.

MOTS-CLÉS/KEYWORDS

segmentation des mots en chinois, traduction automatique, caractère pragmatique de la langue, information contextuelle, modèle cognitif
Chinese word segmentation, machine translation, pragmatically-oriented language, contextual information, cognitive model

1. Status Quo of CWS Systems in MT

For several thousand years, the Chinese language, unlike English and other Western languages, has been written in a continuous string of characters without word delimiters such as white spaces, which presents itself as a unique problem in Machine Translation (MT): how to segment words in Chinese? According to Wu (2008: 631-632), most of the current MT programs adopt either a linguistically-oriented or a statistically-oriented Chinese word segmentation (CWS) system. In a linguistically-oriented CWS system, we first establish a large lexicon that contains (almost) all the

possible words in Chinese, and then apply certain linguistic rules to divide the input sentence into small chunks, which are to be compared with the items (i.e., possible words) in the lexicon. And in a statistically-oriented CWS system, we usually employ word frequency and character co-occurrence probability to determine the word boundaries, with the statistics (such as word frequency and character co-occurrence probability) automatically derived from examples by the system itself. Recent development in the field also witnesses a number of hybrid methods, attempting to garner the strengths of both approaches. (For a more detailed literature review of CWS techniques, see Emerson 2000; Mao, Cheng *et al.* 2007; Wu 2008.) The difference between the above mentioned 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 rules in accordance with their knowledge, expertise, and intuition (Carl, Iomdin *et al.* 2000: 223-224) while inductive systems depend on examples, which usually take the form of a corpus. 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 multiple senses, with a single character capable of forming words with many other different characters (and sometimes with itself), preceding or following 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 correct 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, Gao *et al.* 2003). Although some CWS systems claim to have achieved a success rate of more than 95% in theory, they seldom perform so well in practice. Furthermore, MT is particularly CWS-intensive, because MT systems usually take a sentence as the unit of translation and as long as there is a single CWS mistake in a sentence, the subsequent translation of the whole sentence will be incorrect. In other words, there is an "amplification" effect of CWS errors in MT:

> Suppose that a CWS system has an error rate of 2%. Then, there would be approximately 20 CWS mistakes in a 1000-word article. Let's have another supposition that the average sentence length is 5 words, and then the whole paper contains about 200 sentences. When the CWS mistakes are randomly scattered (not occurring together) in the article, these 20 CWS errors would result in about 20 incorrectly translated sentences, an error rate of 10%. In other words, the rate of CWS mistakes will be "amplified" in the process of translation, and the amplification coefficient approximates the average sentence length. This has an enormous impact upon the success rate of MT. (Liu and Yu 1998: 508-509; translated by the author.)

In an experiment on two MT programs and an automatic word-segmentation program, Wu (2008: 642) finds that the CWS systems in these programs usually could not vary their CWS of a certain linguistic chunk appropriately according to the changed linguistic contexts ("ambiguity" word-segmentation). It should be reiterated that the linguistic chunks under consideration are not really ambiguous. 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 denotes *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, these three programs are quite "consistent" in their CWS of these linguistic chunks, with one never changing its segmentation in all sentence pairs, one making CWS modification in 1 sentence pair, and another adjusting its segmentation in 4 out of the 12 sentence pairs under investigation. Their correct rates of ambiguity CWS are 8.3%, 54.2%, and 66.7% respectively, and if their ability to take adaptive segmentation for the same linguistic chunk in different contexts is taken into account, the correct rates of ambiguity segmentation are 0%, 9.1% and 36.3%. Moreover, wrong word segmentation of a sentence necessarily results in incorrect translation of it. The translation mistakes of these linguistic chunks caused by incorrect CWS make up 78.6% (11/14) and 66.7% (8/12) of all the translation mistakes incurred by these two MT programs. It is little wonder that CWS is often referred to as the bottleneck for Chinese-English MT.

2. Pragmatically-oriented Feature of Chinese

Some studies in contrastive linguistics have shown that European languages are mainly syntactically-oriented while Chinese is basically pragmatically-oriented. After careful comparison between Chinese and some Western languages, Xu Tongqiang (1997: 52), a leading Chinese linguist, argues that the Chinese language is semantically-oriented while Indo-European languages are grammatically-based. In an article, Robertson (2000: 169) refers to Chinese as a "discourse-oriented" language and English as a "syntax-oriented" language. Huang (2000) also makes the claim very explicitly that Chinese is pragmatically-oriented. In other words, semantic and pragmatic consideration plays an important role in understanding Chinese texts, hence significant in CWS. All the above-mentioned CWS systems, however, are either linguistically-oriented or statistically-oriented. Both types have some innate defects that cannot be overcome due to the pragmatically-oriented feature of the Chinese language. In this part, we will go a step further to elaborate on the argument that the Chinese language is basically pragmatically-oriented rather than syntactically-oriented, especially compared with English (Other European languages, such as German and Russian, have more rigorous morphology and syntax than English. However, due to my limited knowledge of these languages, they are not included in this research.).

First, the subject and the object of a Chinese sentence are relative to some extent. For example:

(1)	十个人	吃	一锅饭。	vs.	一锅饭	吃	十个人。
	Ten-people	eat	a-pan-of-rice.	vs.	A-pan-of-rice	eat	ten-people.[1]
(2)	两个人	骑	一匹马。	vs.	一匹马	骑	两个人。
	Two-people	ride	one-horse.	vs.	One-horse	ride	two-people.

The two sentences in each pair may be said to express essentially the same situation. One may wonder how they can keep their meaning constant while the subjects and objects have exchanged their places. Some foreigners are even shocked by such a phenomenon. However, Chinese people approach these sentences from a different perspective, i.e., the relation between subject and object in a sentence is

partly determined by the possible plausible interpretation of the situation, rather than by the distributional features of things/persons only. It is in this sense that we say "Meaning is language in use."

Secondly, the relation between a verb and its object is rather unpredictable. It, however, becomes clear as soon as the verb and its object are put into a larger context, such as a text or a sentence. In other words, the context plays a crucial role in determining the meaning of a Chinese word or phrase. Take the phrases 吃食堂 and 打扫卫生 for instance.

(3)	我们	今天	中午	吃	食堂。
	We	today	noon	eat	canteen.
(4)	小明	忘记了	打扫		卫生。
	Xiaoming	forgot	sweep-and-clean-up		cleanness.

食堂 and 卫生 occupy the object position of the verbs 吃 and 打扫 respectively and it can be argued that they are the objects. However, 食堂 is definitely uneatable; rather, it is the place for the act of eating to take place. Similarly, 卫生 is not the target for the act 打扫. Instead, 卫生 can only serve as the purpose of such an act (卫生 can serve as the result of 打扫, but this interpretation seems inappropriate here.). These two extreme examples illustrate that sometimes the verb-object relation in Chinese can be very strange and unpredictable, deviating far from the standard action-target relation. In fact, the verb-object relation is partly determined by what are the verb and the object and what accompanies them. To put it another way, the verb-object relation is the result of "negotiation" of the verb, the object and their linguistic environment.

Thirdly, Chinese proper names, such as human names and geographical names, have no formal marks, while their English counterparts have their first letter capitalized, a very helpful formal mark for deciphering the proper names. As for personal names, Chinese people tend to give their newborn babies a unique name with a unique meaning. Chinese personal names, with a potentially unlimited total number and no capitalized first letters, make the personal name identifying process of a computer extremely difficult. "[I]t is often difficult, or impossible, to determine when a sequence of characters is being used as a name, and when it is not. This form of ambiguity is difficult to solve without deeper contextual (viz. semantic) information" (Emerson 2000: 9). Concerning the geographical names, we have the same problem of no explicit formal marks such as capitalized first letters. To make the situation even more complicated, both Chinese personal and geographical names are mixed with other words in texts, making it particularly difficult for machines to identify where their boundaries are.

In fact, when we try to figure out the story behind word segmentation in the Chinese language, we find that segmentation itself provides telling evidence for the statement that the Chinese language is structurally weak. There is no such word segmentation problem existing for English and other European languages! These languages usually have a well-developed morphological and syntactical system, and the formal lexis marker, white space, is just part of the system.

The above-mentioned evidence concerning the pragmatically-oriented attribute of Chinese is abundant in the language. Based on these facts, some scholars argue that the structural categories, such as subject and object, cannot account for the

Chinese language sufficiently. Some linguists even go to the extreme of proposing that these structural categories do not exist in Chinese. In their opinion, a Chinese sentence is better analyzed in terms of theme and rheme, with the former occupying the initial position of a sentence and serving as a topic and the latter forming the remaining part and describing or explaining the topic.

Although we do not hold such an extreme opinion, we do agree that Chinese is more pragmatically-oriented than structurally-oriented. Every language expresses meaning in linguistic forms, and Chinese is no exception. But we should be cautious of the relationship between meaning and form. The meaning determines the form, rather than the other way around. The meaning takes precedence. However, some languages, such as most of the European languages, have a well-developed formal system. In these languages, a certain kind of meaning usually corresponds to a particular structure. Therefore, structural information, such as distribution, can be reversely used to help pin down the exact meaning of a word or phrase in a text. However, the formal approach, as the above examples and elaborations have shown, does not work very well in the Chinese language. Structural information only provides a partial picture of the process for detecting meaning of the Chinese language, playing merely a limited role in the process by which humans identify meaning.

As the above analysis shows, the Chinese language is morphologically and syntactically weak, especially compared with English and other European languages. This feature of Chinese subsequently contributes to the pragmatically-oriented feature of the language: if the morphological and syntactical system is not well-developed enough to differentiate meanings, then meaning generation and determination have to be negotiated by all the components. As Chinese is a pragmatically-oriented language, it is no wonder that the available linguistically-oriented and statistically-oriented CWS systems in MT can not successfully deal with the language.

3. A Cognitive Model of CWS for MT

In the previous section, we examined the pragmatically-oriented feature of the Chinese language, which explains why the existing linguistically or statistically-oriented CWS systems in MT are not successful in dealing with the Chinese language. It means that we need to find an entirely new approach to the CWS problem in MT, in which semantic and pragmatic information would be considered or even highlighted. Wu (2008) has conducted a language investigation consisting of two surveys and eight interviews and probed into the question how Chinese people segment a Chinese sentence into words in their reading, with the hope that the knowledge of human's word-segmentation process may shed light on CWS techniques. The results obtained are as follows. Human beings achieve a relatively homogeneous word-segmentation result, obtaining an 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 and pragmatics unconsidered, which largely accounts for their poor word-segmentation performance. Additionally, CWS systems in MT are usually independent of

other processing stages, wasting part of the information obtained from the CWS process.

Based on these findings, some implications are drawn in the hope that they can be used in MT system designing, especially regarding the part of CWS. First and foremost, we could incorporate into word-segmentation systems contextual information to improve their performance. Second, semantics should play a more important role in the CWS system. Third, since the word segmentation is a dynamic process and cannot be settled in a single pass, the one-way mechanical processing approach may need to be discarded in favor of a bi-directional one.

Last but not least, as CWS is only part of the natural language understanding process, we might retain some of the processing information during CWS for later use, especially during the meaning determination and target word selection stages. This means we'd better not standardize the CWS processing. Instead, we could standardize some smaller components of CWS (such as the lexicon and the knowledge-based word-segmentation rules), allowing the CWS system to interact with other stages of MT or to be used interactively in other linguistic and translation tools. (For further discussion of the advantages of standardized CWS components, see Wu 2008: 645-646.) Besides, we intend to abide by the LISA SRX (Segmentation Rules Exchange) format, TBX (Termbase Exchange) format, TMX (Translation Memory Exchange) format, ISO CD24614-1 (Language Resource Management), and other standardization guidelines in the programming, so that the CWS results, as well as the lexicon and other CWS components, could be readily used in other standardized linguistic and translation tools, such as Information Retrieval (IR) systems, Translation Memory (TM) managers, Terminology Database (TD), part-of-speech (POS) taggers and other MT programs.

These implications, we think, deserve our attention, which might be the possible ways to improve the current CWS techniques. In light of the above-mentioned knowledge of human CWS process, we have built a new word-segmentation model (Figure 1), aiming to address the word-segmentation problem in machine translation from a cognitive perspective. In this section, we will explain in detail how this newly-designed model works.

3.1 Preliminary Processing Stage

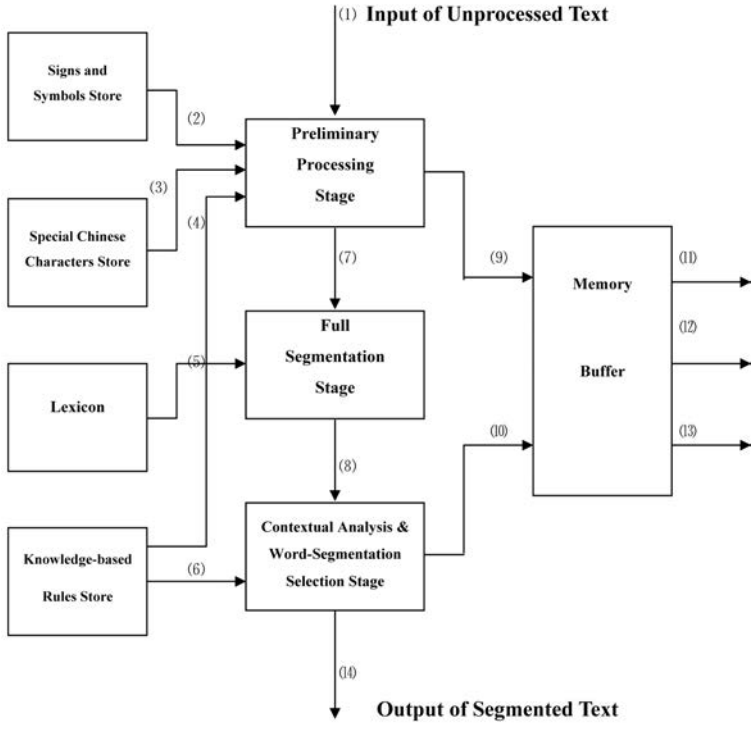
The new model consists of three main stages: Preliminary Processing, Full Segmentation, and Contextual Analysis and CWS Selection (Figure 1). We shall introduce the functions of these stages one by one.

The Preliminary Processing Stage, implied by its name, is responsible for the word-segmentation preprocessing. It intends to accomplish the following two tasks: 1. to input sentences for CWS and deal with signs and symbols (see (2) (4) in Figure 1); 2. to pick up special Chinese characters in the text (see (3) (4) in Figure 1).

3.1.1 Sign and Symbol Processing

First, the Preliminary Processing acts as the initial threshold to admit a suitable chunk of the text waiting in the CWS processing queue. Usually, this suitable chunk is a sentence, denoted by a string between two neighboring sentential punctuation marks (full stops, question marks, or exclamation marks). Although this stage of Sign

FIGURE 1
Word Segmentation Processing Engine



and Symbol Processing usually focuses on one sentence, the system also brings in two other sentences preceding and following the focused one (except at the beginning or the end of a text), so as to provide enough contexts for word segmentation.

Apart from the three sentential punctuation marks mentioned above, there are other punctuation marks in Signs and Symbols Store,[2] such as comma, colon, semicolon, quotation marks, Chinese book title mark (《》), traditional Chinese quotation mark (『』). Although these punctuation marks are merely clues for a fragment of a sentence (rather than a sentence), they indicate a kind of natural division of different parts of a sentence and, as a result, are useful for CWS.

This system component also handles other signs and symbols that are in Signs and Symbols Store. Among them are numbers, foreign letters, etc. Numbers are tackled with the help of knowledge-based rules (see (4) in Figure 1). First, ordinal numbers, by these rules, are separated from cardinal numbers. Then, a number is handled by checking its collocation, and segmenting it together with the quantifier or the unit word following it if they form a kind of 'numeral + quantifier' or 'numeral + unit' relation. Besides, 'numeral + quantifier' or 'numeral + unit' collocations will be further classified by knowledge-based rules. For example, temporal phrases, such as 'numeral + 月 (*month*),' will be selected and sent to the subsequent translation stage for tense selection. (e.g., 5月 (literally *five+month*) would be selected and later rendered as *May* rather than *five months*.)

Foreign letters, including alphabetic letters, words, phrases and sentences, are examined by these knowledge-based rules too. The examination aims to see if these foreign letters are used alone or together with Chinese characters (e.g., T恤衫, meaning *T-shirt*). These foreign letters that are used alone are picked up and segmented while those 'foreign letter(s) + Chinese character(s)' are segmented and then translated as an organic whole. For example, T恤衫 would be counted as a word and later translated into *T-shirt(s)*.

Apart from these results that will undergo Special Chinese Character Processing, the information concerning judgments and decision-making by knowledge-based rules will also be recorded and then sent to Memory Buffer, a storehouse for temporary files via flow line (9) in Figure 1. This information is to be used in the later translation stage, such as meaning determination and target word(s) selection (see (11) (12) in Figure 1). For example, based on the results of Sign and Symbol Processing, 'numeral + 月 (*month*)' (as compared to 'numeral + 个 +月(*month*)') will be translated as a certain month in a year rather than 'numeral + months' (i.e., several months) in English whereas T恤衫 will be rendered as a word (i.e., *T-shirt*) rather than several English words.

The CWS in this stage is called "gold segmentation." The segmentation is 'gold' because in most cases it will not be changed and adjusted during later processing stages. The results from this stage will be retained for further processing, i.e., during the Special Chinese Character Processing stage.

3.1.2 Special Chinese Character Processing

Special Chinese Character Processing is accomplished by collaboration of Special Chinese Characters Store and Knowledge-based Rules Store (see (3) (4) in Figure 1).

First, the system applies Special Chinese Characters Store in the selection of special Chinese characters in the text. These special Chinese characters mainly consist of auxiliary words for tense, such as 了, 过, 着 and some other special characters like words with affixes (e.g., 老虎, 老鼠), and words formed by doubling the same character(s) (e.g., 明明白白, 调整调整).

Then, Knowledge-based Rules Store is used to differentiate these special Chinese characters from other words. This process can be illustrated by looking at an example of differentiating auxiliary words for tense from non-auxiliary words. In the sentence 我 十分 了解 他 [literally: *I very-well know him*], 了 forms a word 了解 with the character 解. So it is not a tense auxiliary. These non-auxiliary words will be filtered out. In the same vein, words with affixes and words formed by doubling the same character(s) will be checked and separated from other parts of the text in this processing stage.

The results of processing in this stage will be recorded and sent to Memory Buffer (see (9) in Figure 1) as well as to the subsequent processing stage, Full Segmentation (see (7) in Figure 1). The information stored in Memory Buffer will subsequently be employed in the later translating process, especially for the tense-selection stage (see (13) in Figure 1).

3.2 Full Segmentation Stage

After the preliminary processing, the text goes to the next processing stage, Full Segmentation.

During this stage, Lexicon, a dictionary-like store that contains (almost) all possible words in Chinese, is applied to performing word segmentation on the text. In the existing CWS systems for MT, certain structural and grammatical rules are applied to divide the input sentence into small chunks, which are to be compared with the items (i.e., possible words) in the lexicon. This approach can be divided into subcategories, such as Maximum Matching and Minimum Matching on the one hand, and Obverse Matching and Reverse Matching on the other. The first distinction is based on the criterion of giving priority to either long words or short words, and the latter pair of matching strategies varies with respect to the direction of processing. Among these approaches, the Maximum Reverse Matching Method has been most widely used. This method, however, is notorious for its poor treatment of ambiguity processing of CWS (Liu 2000; Wang, Gao *et al.* 2003; Luo, Chen *et al.* 1997; Yin 1998). Different from the existing word segmentation programs, the text will undergo four kinds of word-segmentation in this new model: Maximum Obverse Matching, Minimum Obverse Matching, Maximum Reverse Matching, and Minimum Reverse Matching. In this way, all the possible words are counted and recorded, hence the name *Full Segmentation*. Of course, full segmentation will increase computational overhead to some extent and thus reduce algorithmic efficiency, but it surely improves the quality of CWS. Given that algorithmic capacity of CPU is being updated and enhanced day by day, it is possible for us to employ the technique of full segmentation without drastically compromising the CWS speed.

The results of the full segmentation stage are sent to the subsequent processing stage Contextual Analysis and CWS Selection, which is the most important stage in this model (see (8) in Figure 1).

3.3 Contextual Semantic Analysis and Word Segmentation Selection Stage

Once the text has received full segmentation, it will be sent to the main processing stage "Contextual Analysis and Word Segmentation Selection." In this stage, the text undergoes contextual semantic analysis by checking against the rules from Knowledge-based Rules Store (see (6) in Figure 1). Subsequently, an optimal word-segmentation is established by choosing and restructuring the correctly segmented words from all the possible words derived from Full Segmentation. Before going into the details of the processing chart of this stage, we should look at what a knowledge-based rule is like, the knowledge of which will render the depiction of the processing procedure more comprehensible.

3.3.1 Knowledge-based Rules

A knowledge-based rule,[3] like an entry word in a dictionary, is made up of several items, which, in turn, consists of two layers. A single rule usually deals with one word, whose different senses become the different items of the rule. Each item of the rule has two layers, that is, the explication of the sense and the context(s) that the word typically or usually goes with when it takes this sense. The explication of sense may encompass several kinds of semantic interpretation apart from the usual semantic

explanation, such as semantic components, synonyms, antonyms, superordinates, hyponyms. The context(s) that the word typically or usually goes with when it takes this sense is in fact referred to the usage(s) of the word when it denotes this sense. However, the usage(s) is substantiated by specifying the context(s), collocation(s), semantic link(s), and structural pattern(s) it typically or usually goes with.

It should be noted that the number of the items of a certain word also tallies with the number of the word's senses and target words in the later translating stage. To be more specific, if a word has five different senses, the knowledge-based rule for processing it will have five items, and the semantic items for this word in the meaning determination stage will also be five. It is also true of the number of target words for rendering this word, which will correspond to its five senses. Therefore, if we can decide which sense the word takes in the present context, it will be of great help for meaning determination and target word(s) selection in the later translating stage.

Besides, it is worth mentioning that these knowledge-based rules are expected to be written from a cognitive perspective. A knowledge-based rule, just as depicted in the previous two paragraphs, merely deals with a single word, the senses of which are listed, not according to the frequency of each sense, but according to the order at which each sense comes to an ideal speaker's mind when he/she encounters the word in a certain context (i.e., "salience" of each sense of the word, or "ease of activation" in Langacker's term), though the more salient sense tends to be the more frequent one (see Taylor 2002: 123-139; Langacker 1987:34-40). Consequently, the different senses of the word are ordered according to the salience of each, with the most salient one as the default value. However, the salience of a sense is relative: the original balance could be broken by the appearance of a certain contextual condition, with a less salient sense strengthened and becoming more salient.

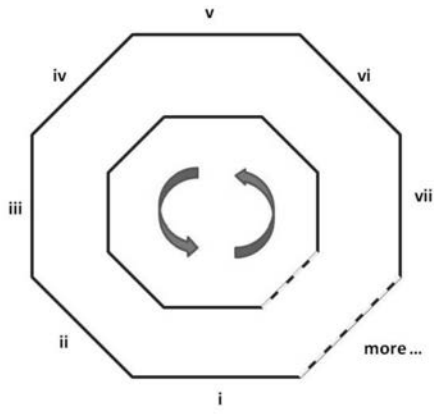
3.3.2 Contextual and Semantic Analysis and Word Segmentation Selection

Now that we have established how knowledge-based rules work, we can examine the detailed procedures of the processing stage.

First, the Contextual Analysis and Word Segmentation Selection Stage receives the input from Full Segmentation stage via flow line (8) (Figure 1). The input consists of three sentences, carrying all the information obtained from the previous two stages Preliminary Processing and Full Segmentation. Although there are altogether three sentences under processing, the focus is just on one sentence. The other two sentences are the sentences that either precede or follow the sentence, and they are expected to provide enough contexts for the sentence under segmentation. In other words, a single sentence remains in this processing stage and is actually subjected to three different processing passes: first, it is input so as to supply contextual information for processing the sentence preceding it; then, it undergoes processing itself; after it has been processed, it will still be retained as contextual clues for segmenting the following sentence. Although in three rounds of processing a sentence receives the spotlight only once, its processing result is not merely determined by this round of processing. Rather, the result of processing this sentence could be modified or changed in any one of the three rounds of processing. Therefore, we think that the processing unit of the system is three sentences.

Then, the input ignites the processing engine. The engine can be visualized as two homocentric polygons (Figure 2).

FIGURE 2
Word Segmentation Model for Machine Translation



The engine equipped with a knowledge-based rule can only deal with one single word, with each lateral representing one of the items (one of the word's senses) and each polygon symbolizing one layer of the items (the inner polygon for the explication of senses and the outer polygon for the usual or typical contexts concomitant with these senses). The two polygons are homocentric and have the same number of laterals, with each inner lateral corresponding to one outer lateral. That is, the explication of a sense of the word (the inner lateral) corresponds to the context(s) that the word typically or usually goes with when it takes this sense (the outer lateral). Each pair of laterals, as a result, represents one of the senses and its usages of a certain word, and all the laterals together make up a rule-equipped polygon engine, which encompass all the senses and usages that the word has. As a result, the number of pairs of laterals for the two polygons is variable, usually ranging from 3 to 20. (If a word only has one or two senses, each of the two polygons, rather than becoming one or two lines, will still have 3 laterals with two or one lateral getting an empty value.) When a prospective word X is fed into the processing engine, the knowledge-based rule of the word X comes in from Knowledge-based Rules Store and arms the engine, with each sense of X becoming a pair of laterals of the polygon engine. Then the engine starts the anti-clockwise processing sequence (as the arrows in Figure 2 show) and begins to search among all the possible words for the context that a usage of A typically or usually goes with. If a prospective word Y from Full Segmentation matches an item (representing a usage of the word X), the corresponding lateral pair will be weighted, and when the engine stops, both words (X and Y) become salient, with X picked up as a segmented word and the word Y marked. At this time, the engine is ready for processing another word.

However, if there are no words matching any lateral, the engine starts matching the inner polygon with all the possible words, and carries out a similar process. If this round of matching still does not work, the engine will take Lateral Pair I, the default value of the engine, as the solution and proceed to process another word.

There is another situation that starts the inner polygon. If the context of the word contains 是, 也就是(说), 等于(说), (的)意思是, 定义(是), and other meta-language

(Some of their English counterparts are *that is*, *i.e.*, *in other words*, *to be defined as*.), then the inner polygon will be triggered and carry out a matching process.

Of course, it is possible for a lateral to gain weight several times or for several laterals to be weighted. If this is the case, the summation and comparison of weight will be carried out. In the case that several laterals have been weighted and there exists conflicting word-segmentation, the lateral pair that gains more weight will become the most salient item and be picked up by the system.

The matching process is followed by the stage of selection and reconstruction. The system will first pick up the salient words that have yielded one or more successful matches, and then fill out the other parts of the sentence by selecting the default words. The segmented text will be output via flow line (14).

This stage of system has another function: it performs another processing on temporal words and word groups. First, the words or word groups denoting time are distinguished and singled out. Then two kinds of time, i.e., a point of time (e.g., ten years ago), and a period of time (e.g., for ten years), are differentiated and recorded. The information obtained will also be stored in Memory Buffer and then be sent to the tense selection stage (see [13] in Figure 1).

As far as the distribution of the processing results is concerned, there is still another processing stage that feeds results into Memory Buffer. The results from contextual and semantic analysis (e.g., the sense a word takes in the context, the collocation the word forms with other words), which will be first stored in Memory Buffer (see (8) (10) in Figure 1) and then be sent to the translating stage and used to determine the sense and select target word(s)/structure(s) (see (11) (12) in Figure 1).

Since both the output of segmented text (from Route (14)) and the information stored in Memory Buffer (from paths (11), (12) and (13)) are sent to the MT stage, there arises a question – What's the difference between these two kinds of information? The output of flow line (14) is the segmented text. It is the finished product (therefore it could be directly used in other related Natural Language Processing areas), but does not carry much information about the decision-making process. The information stored in Memory Buffer, however, is mainly concerned with the decision-making process. In other words, if we say the Output of Segmented Text tells us which is a word, the information stored in Memory Buffer tells us why it is a word. If these two kinds of information are considered together, we may have a better understanding of the text, and therefore improve the performance of MT in the later stages.

4. Concluding Remarks

MT is not just a matter of technology. It also entails insights and expertise from linguists. This article is part of the research which investigates how human beings carry out Chinese word segmentation by a language investigation consisting of two surveys and eight interviews, and then draws inspiration from the cognitive process of human's CWS and teaches machines how to do word segmentation. Anyway, why can't we learn something from our fellow homo sapiens? It is mainly in this sense that we call this research 'a cognitive perspective.' Compared with the currently available CWS systems of MT, this model has incorporated some elements of humans' unique CWS mechanism and improved the CWS techniques in the following ways: 1) pragmatic and contextual information has been utilized, which is enlarged to

include three sentences rather than restricted to immediate context; 2) Semantics plays a more important part in the CWS system, which is encapsulated in the knowledge-based rules and embodied by the senses and usages of a word in the polygon engine; 3) The one-way CWS approach is discarded and a bi-directional one is employed; 4) Not only the segmented text, but also some other useful information concerning the decision-making process in the CWS system, have been retained for later use, especially in the stages of meaning selection and target word determination. In this way, the CWS program could interact with other processing stages in MT, and different stages of the macro MT system might achieve consistency in their analyzing of the text under consideration. Hopefully, the new CWS model will not only enhance the CWS performance, but also improve the translation quality of MT systems.

NOTES

* The research has been supported by Model Postgraduate Course Project of Jiangsu Province of China (江苏省优秀研究生课程，苏教研［2010］6号) and by the NUST Research Funding, NO. 2011YBXM136.

1. All examples are translated by the author of the article if not otherwise stated. To make examples and their translations more comprehensible, some white spaces are inserted by the author between words and phrases in Chinese. The English counterparts of these words and phrases are hyphened to make the correspondence more explicit. For example, the phrase 十个人 is rendered here as *ten-people* rather than *ten people*.
2. *Store,* rather than *lexicon*, is employed here for the reason that *Signs and Symbols Store* contains punctuation marks, numbers, foreign letters, etc. These signs and symbols cannot be classified as Chinese words. This is also the case for *Special Chinese Characters Store*. It contains prefixes, suffixes, auxiliaries and so on, which do not sit easily in the category of Chinese words in the narrow sense. Therefore, the use of *lexicon* seems not very suitable.
3. A distinction has been made between the two terms *knowledge-based* and *rule-based*. *Rule* in the term *rule-based* usually refers to the linguistic rules, especially morphological and syntactical rules, in most cases employed in a linguistically-oriented CWS system. Just as we have pointed out, grammatical rules are not very effective in CWS systems. By *knowledge-based* rules, we mean semantic and pragmatic information should also be incorporated into the process of CWS.

REFERENCES

Carl, Michael, Iomdin, Leonid and Streiter, Oliver (2000): Towards a Dynamic Linkage of Example-based and Rule-based Machine Translation. *Machine Translation*. 15(3):223-257.

Emerson, Thomas (2000): Segmenting Chinese in Unicode. *In*: *Proceedings of the 16th International Unicode Conference* (Amsterdam, 27-30 March 2000). Visited on 15 August 2010, <http://seba.ulyssis.org/thesis/papers/iuc16.pdf>.

Huang, Yan (2000): *Anaphora: A Cross-linguistic Study*. Oxford: Oxford University Press.

Langacker, Ronald (1987): *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford: Stanford University Press.

Liu, Kaiying (2000): *Automatic Word-segmentation and Tagging for Chinese Texts* (中文文本自动分词和标注, Zhong wen wen ben zi dong fen ci he biao zhu). Beijing: The Commercial Press.

Liu, Qun and Yu, Shiwen (1998): Difficulties in Chinese-English Machine Translation. *In*: Changning Huang, ed. *Proceedings of the 1998 International Conference on Chinese Information Processing* (1998中文信息处理国际会议论文集, 1998 Zhong wen xin xi chu li guo ji hui yi lun wen ji). Beijing: Tsinghua University Press, 507-514.

Luo, Zhengqing, Chen, Zengwu, Wang, Zebing, *et al.* (1997): A Review of the Study of Chinese Automatic Segmentation (汉语自动分词研究综述, Han yu zi dong fen ci yan jiu zong shu). *Journal of Zhejiang University*. 31(3):306-312.

Mao, Jun, Cheng, Gang, He, Yanxiang, *et al.* (2007): A Trigram Statistical Language Model Algorithm for Chinese Word Segmentation. *In*: Franco P. Preparata and Qizhi Fang, eds. *Frontiers in Algorithmics*. Berlin/Heidelberg: Springer, 271-280.

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

Taylor, John (2002): *Cognitive Grammar.* Oxford: Oxford University Press.

Wang, Ke, Gao, Changbo, Zhai, Xuefeng, *et al.* (2003): The Main Techniques in Chinese Word-segmentation and Its Prospect of Application (汉语分词的主要技术及其应用展望, Han yu fen ci de zhu yao ji shu ji qi ying yong zhan wang). *Communications Technology*. 138(6):12-15.

Wu, Zhijie (2008): New Light Shed on Chinese Word Segmentation in MT by a Language Investigation. *Meta*. 53(3):630-647.

Xu, Tongqiang (1997): *On Language: Structural Rules and Research Methodology of the Semantically-oriented Language* (语言论—语义型语言的结构原理和研究方法, Yu yan lun – yu yi xing yu yan de jie gou yuan li he yan jiu fang fa). Changchun: Northeast Normal University Press.

Yin, Jianping (1998): Automatic Word Segmentation Methods for Chinese Language (汉语自动分词方法, Han yu zi dong fen ci fang fa). *Computer Engineering and Science*. 20(3):60-66.