自然言語処理 123－3
（1998. 1. 19）

The Concept of Sensitive Word in Chinese Machine Translation

Fuji Ren
Faculty of Information Sciences, Hiroshima City University
Japan, ren@its.hiroshima-cu.ac.jp

Jian-Yun Nie
Department of Computer Science and Operation Research, University of Montreal
Canada, nie@iro.umontreal.ca

Abstract In Machine Translation (MT), using compound words or phrases makes the translation process easier. For example, the phrase "信息高速公 " corresponds unambiguously to "information highway". It is not necessary to break it down to "information" and "highway". However, some compound words (phrases) in Chinese are composed of simpler words which can play significantly different roles in sentences when they are broken down. For example, the compound word "机器翻译" (machine translation) may be broken into "机器" (machine) and "翻译" (translate), as in the sentence "他用机器翻译文 " (He uses a machine to translate papers). We call such a compound word a Sensitive Word. During Chinese MT processing, if the first segmentation result leads to a failure, the alternative solution with a sensitive word broken down is considered as the preferred one. This allows us to reach at a higher efficiency by avoiding examining impossible segmentation solutions. In this paper, we describe the problems related to sensitive words. A machine readable dictionary has been examined, and more than 800 sensitive words have been found. This shows that sensitive word is a common phenomenon in Chinese that is worth closer examination.

Keywords Machine Translation, Sensitive Word, Chinese, Compound Word

中国語機械翻訳における敏感語概念

任 福継* Jian-Yun Nie**
*広島市立大学情報科学部
ren@its.hiroshima-cu.ac.jp
**モントリオール大学計算機科学部
nie@iro.umontreal.ca

あらまし 機械翻訳を代表とする自然言語処理において、文の解析を高精度、かつ効率良く行うという観点からすれば、できるだけ複合語単位で処理することが望ましい。例えば、中国語”信息高速公路”（ハイウェー）について我々は”信息”と”高速公路”に分割する必要がない。しかし、中国語の複合語は、場合によっては分割しなければ正しく解析できないことが多い。例えば、”他用机器翻译文章”（彼は機械で文献を翻訳する）中の”机器翻译”について、”机器”（機械）と”翻译”（翻訳＋する）に分解する必要がある。このような複合語を本論文では敏感語と呼ぶことにする。中国語文の処理において、形態素解析のある候補を用い構文解析や意味解析などを処理する途中で失敗した場合、このような敏感語のみを次の解析候補とすれば、すべての複合語を解析候補とする必要がなくなる。それで、敏感語という概念を導入することにより、高効率的な中国語文処理が期待できる。本論文では、上述の敏感語についての概念を提案し、さらに87599語の中国語辞書を用い実験と検討を行った結果を報告する。

キーワード 機械翻訳、敏感語、中国語、複合語

1. Introduction

Natural language processing (e.g. Machine Translation, message understanding) usually starts from words. The identification of words in Indo-European languages is a trivial task. However, this problem has been, and is still (as we will argue) a bottleneck for Chinese. In Chinese, sentences are written as continuous character strings without word separation. For example, "计算机已经用于各个领域 " (Computers have been used in every area).

Much work has been dedicated to segmenting Chinese correctly. It has been often reported that the segmentation accuracy may reach at as high as over 99%. This gives one the impression that the segmentation problem has been successfully solved. However, the tests for segmentation have been oversimplified: segmentation has been tested as if it is the finality of Chinese processing. If segmentation is considered together with further processing (e.g. tagging), many hidden problems will be unveiled. In addition, the test corpus was often built by the same people who programmed the segmenter. So there may be a bias in performance measurement. If the segmentation problem is considered together with the use we will make of it (e.g. for machine translation), the accuracy rate is much lower. This fact has been confirmed by a recent survey [11] conducted by a group of experts in China. They used several test corpora for different tests. This survey gave the following conclusions about current state of Chinese NLP software:

1) On the test for word segmentation and tagging:
   The best software reached at an 89.4% accuracy for word segmentation, and only 79.58% of words have been tagged correctly in the best case.
2) On word boundary determination:
   The best success rate for word boundary determination is 60%, and the global accuracy is 89.4% (1). The accuracy rate for tagging correctly segmented texts is 86.6%.
3) On the test for solving ambiguities in word segmentation:
   The systems had success rates varying from 61% to 78% for dealing with overlapping ambiguities (see next section for a description), and from 36% to 59% for dealing with combinatory ambiguities (see next section).

These conclusions make it clear that the claimed over 99% accuracy rate hid many unsolved problems. These problems will greatly affect the quality of further processing on Chinese.

In this paper, we intend to deal with one of the segmentation problems - the combinatory ambiguity which is the most difficult to solve. Our goal is not to determine the correct segmentation. This is extremely difficult to achieve. Instead, in case the first segmentation solution leads to a failure, we try to identify the next best segmentation solution which indeed is the most likely solution. To do this, we have to determine which word in a sentence raises combinatory ambiguity. Not all words that can be separated into simpler words raise combinatory ambiguities. They do only when the component words, when separated, may play significantly different roles than their compound in sentences. We call such words *sensitive words*. A sensitive word will be considered as a preferred breaking point when alternative segmentation is sought.

In the remaining of the paper, we will first describe the segmentation problem and the impact of sensitive words on it. Then we will report the result of our preliminary examination of a machine readable Chinese dictionary which unveils a great number of sensitive words in it.

2. Ambiguities in segmentation

Let us show some examples first.

Example 1: 王老大将来东京读书.

---

(1) Accuracy for word boundary $= 1 - \frac{\#\text{ errors}}{\#\text{ boundaries}}$, Global accuracy $= 1 - \frac{\#\text{ errors}}{\#\text{ words} / p}$

where p is the average word length (=1.62).

(Wang Laoda will come to study at Tokyo)

This sentence may be segmented into several legitimate words. Two of them are as follows:

Segmentation 1.1: 王老大 / 将 / 来 / 东京 / 读书 / 。
Wang Laoda / will / come / Tokyo / study /.

Segmentation 1.2: 王老大 / 将来 / 东京 / 读书 / 。
Wang Laoda / in the future / Tokyo / study /.

In this example, the string 将来 may be segmented in two different ways: as 将来 - in the future, or as 将/来 - will come. Thus, there is a segmentation ambiguity. The first segmentation is the correct one whereas the second one is not grammatical.

Example 2: 他研究生物化学。
(He studies biochemistry)

Two possible segmentation solutions are as follows:

Segmentation 2.1: 他 / 研究 / 生物化学 / 。
he / study / biochemistry / .

Segmentation 2.2: 他 / 研究生 / 物 / 化学 / 。
he / graduate student / object / chemistry /.

The character 生 may be grouped either with the precedent characters 研究 to form 研究生 (graduate student) or with the next characters 物化学 to form 生物化学 (biochemistry).

Usually, segmentation ambiguities may be separated into combinatory ambiguity and overlapping ambiguity, according to their position configuration. Examples 1 and 2 show these cases respectively: the words 将 (will) and 来 (come) may be *combined* to form another word 将来 (in the future) in the first example; the words 研究生 (graduate student) and 生物化学 (biochemistry) *overlap* in the second example.

We can also separate segmentation ambiguities according to their nature (morphological, syntactic and semantic).

- Morphological ambiguity

Both example shown above contain a morphological ambiguity. This kind of ambiguity may be solved by using syntactic analysis. In the two examples, the second incorrect segmentation may be ruled out because they are not grammatical.

- Syntactic ambiguity

Let us look at the following example.

Example 3: 物理学起来很难。
(Physics is difficult to learn)

It may be segmented into two grammatically correct word sequences:

物理 / 学 / 起来 / 很 / 难 / 。
(Physics is difficult to learn)

or 物理学 / 起来 / 很 / 难 / 。
(Physics is difficult to stand up)

This kind of ambiguity may only be solved by using the semantic information that tells "stand up" is incompatible with "physics".

- Semantic ambiguity

Example 4: 美国会采取措施支持日本。

This sentence may be interpreted as "The United States will take measures to support Japan",

or "The Congress of the United States will take measures to support Japan". Both interpretations are syntactically and semantically correct. They correspond to the following segmentations:

Segmentation 1: 美国 / 会 / 采取 / 措施 / 支持 / 日本 /.
the U.S.A. / will / take / measure / support / Japan /.

Segmentation 2: 美 / 国会 / 采取 / 措施 / 支持 / 日本 /.
the U.S.A. / congress / take / measure / support / Japan /.

This kind of ambiguity is extremely difficult to solve. It may only be done by analyzing the context in which the sentence is used.

3. Approaches to Chinese segmentation

There have been two main groups of approaches to Chinese segmentation: dictionary-based approaches and statistical approaches.

Dictionary-based (also called rule-based) approaches [2, 5-7, 10] operate according to a very simple concept: a correct segmentation result should consist of legitimate words - the words stored in a dictionary or derivable from some rules. In general, however, several legitimate word sequences may be obtained from a Chinese sentence. The longest matching algorithm is then used to select the word sequence which contains the longest (or equivalently, the fewest) words. For example, in Example 2, Segmentation 2.1 is preferred to Segmentation 2.2 because it contains less words (and longer words). However, this approach fails for Example 1 for which the wrong solution 1.2 will be preferred.

Statistical approaches [1, 3, 4, 8, 9] rely on statistical information such as word and character (co-)occurrence frequencies in the training data - often a set of manually segmented texts. The statistical data describe how probable a character string may be a word, or how probable a word may follow another word. Using these data, the segmentation process consists of first determining the probability of a word sequence, and then choosing the one with the highest probability. For example, in Segmentation 1.2 of Example 2, the character 物 (object) alone is rarely used as a word. So the probability for it to be a word is low. This will make the whole word sequence in Segmentation 2.2 to have a lower probability than Segmentation 2.1. Thus the correct segmentation may be chosen.

Both approaches have difficulty to deal with Example 1. As segmentation 1.2 is shorter than Segmentation 1.1, the longest-matching algorithm will choose Segmentation 1.2. By statistical approaches, as the word 将来 (in the future) is at least as common as 将 (will) and 来 (come) separately, Segmentation 1.2 will also likely have a higher probability than segmentation 1.1. So the segmentation solution suggested first will be wrong, and it will lead to a failure during syntactic analysis. Solving such an ambiguity will require syntactic and even semantic analyses. We rather suggest to tolerate such an error, but to try to reach at the correct solution as quickly as possible. So the problem we raise is: if the first segmentation solution fails, what next solution to propose ? A naive approach would to examine all the best alternatives in turn. For Example 1, the following segmentation solutions will likely be examined if we use the longest matching algorithm:

王 / 老大 / 将来 / 东京 / 读书 /.
王老 / 大将 / 来 / 东京 / 读书 /.
王老 / 大 / 将来 / 东京 / 读书 /.
王老大 / 将 / 来 / 东 京 / 读书 /. (correct)
王老大 / 将来 / 东 / 京 / 读书 /.
王老大 / 将来 / 东京 / 读 / 书 /.

We can see that the correct solution is just one among several others. Before reaching at the correct solution, syntactic (and even semantic) analyses have to be carried out on several other solutions to reject them. The efficiency will be greatly affected.

4. Sensitive words and their impact on MT

If we analyze the words involved in Segmentation 1.1 of Example 1, we can obverse that, although several words may be broken into simpler words, the simpler words do not have equal chances to be correct words in Chinese sentences. For example, it is rare, if not impossible, that the string 东京 (Tokyo) should be separated into words 东 (east) and 京 (capital) in normal Chinese sentences. Words such as 读书 (study) can be separated in a more natural way into 读 (read) and 书 (book). However, they will form together a verb phrase 读书 which has exactly the same meaning as 读书 together. So it is useless to consider such a string as a breaking point.

The word 将来 (in the future) is different from the two last cases. Its components 将 (will) and 来 (come) are both common words in Chinese, and these two words can be naturally put side by side in Chinese sentences. In addition, when they are separated, they play roles significantly different from that of their compound. We call such a word a *sensitive word*. More precisely, a sensitive word is a compound word in which:

- the component words can be naturally put side by side in the same order as in the compound word in normal Chinese sentences;
- and the roles of the component words in such sentences are different from that of the compound.

If a sentence contains a sensitive word, and it has been proposed as a single word in the first segmentation solution that finally failed, then the sensitive word should be first separated into two words. The result of such a separation should be considered as the next best segmentation solution. In Example 1, if the string 将来 is identified as a sensitive word, once Segmentation 1.2 fails, Segmentation 1.1 will be considered as the next best segmentation. This allows us to avoid examining other less likely word sequences.

This approach is perfectly compatible with the common practice in MT to group words into long phrases. The advantage of grouping words into phrases is its high efficiency: if an expression may be unambiguously recognized, it is useless to analyze its composition. For example, it is of no help, and even misleading to analyze the composition of 中华人民共和国 (The People's Republic of China) as follows:

/ 中华 / 人民 / 共和 / 国 /
/ 中华 / 人民 / 共和国 /
/ 中华人民 / 共和国 /
/ 中 / 华人 / 民 / 共和 / 国 /

Long phrases are extensively used in Chinese MT. In fact, many items stored in Chinese dictionaries are indeed phrases. Some of them are sensitive words. It is important to distinguish them from non-sensitive words in order to increase efficiency. The next step of our study is to analyze a machine readable Chinese dictionary to find sensitive words in it. The dictionary contains 87599 items.

The meaning of the symbols we will use is given in Table 1.

Symbol	Meaning	Example
ADJ	adjective	内 (inside)
ADV	adverb	后 (after)
CONJ	conjunction	同 (and)
NC	common noun	机器 (machine)
PREP	preposition	比 (in comparison with)
QUAN-CL	quantity + classifier	几个 (several)
VAUX	auxiliary verb	能 (can)
VPRED	precedent verb	想 (intend to)
VINT	intransitive verb	读书 (study)
VTR	transitive verb	管理 (manage)
V	VINT or VTR	

Table 1. Symbols and their meaning.

In many cases, compound words (phrases) are constructed from simpler words according to

some rules. The following rules are commonly used:

NC + NC —> NC(公司职员 - employee of a company)
NC + V —> V (人工繁殖 - artificially breed)
NC + V —> NC (机器制图 - drawing by machine)
VTR + NC —> NC (进口 - import)
VTR + NC —> VINT (锄草 - weed)
V + V —> V (编印 - edit and print)
ADV + V —> V (自动检测 -automatically detect)
PREP NC —> ADV (如期 - in time)
ADJ NC —> NC (新时代 - new era)

However, there is a complete set of word construction rules, and many exceptions are observed. In any case, sensitive words are not due to the fact that we do not know all the composition rules. Rather they are because Chinese words or characters may often correspond to several categories. A composition may be broken down, and its components separately may play different roles than when they are grouped together. For example, 把门 (guard the door) is composed of 把 (guard) and 门 (door) according to the rule V + NC —> NC. However, 把 is usually used as a preposition. So the same string 把门 may also be considered as a PREP+NC combination, and this combination may be found in ordinary Chinese sentences. So 把门 is a sensitive word.

In our examination of the dictionary, we intuitively determined a set of compositions that are the most propitious for sensitive words according to Chinese grammar. These compositions are shown in Table 2. For example the composition ADJ+VAUX means that the components of a word may also be an adjective and an auxiliary verb. Some examples are given in Table 3 to illustrate each case.

Composition	NB. Words possibly having the composition	NB. Sensitive words (percentage)
ADV + ADV	402	11 (2.7%)
ADV + NC	2231	104 (4.7%)
ADV + PREP	295	10 (3.4%)
ADV + QUAN-CL	10	5 (50.0%)
ADV + VAUX	149	4 (2.7%)
ADV + VINT	1266	21 (1.7%)
ADV + VTR	1156	88 (7.6%)
CONJ + NC	132	15 (11.4%)
CONJ + VTR	95	25 (26.3%)
NC + PREP	629	46 (7.3%)
NC + VAUX	250	96 (38.4%)
NC + VTR	3995	253 (6.3%)
PREP + NC	749	40 (5.3%)
PREP + VTR	454	17 (3.7%)
VPRED + PREP	97	8 (8.2%)
VPRED + VTR	364	22 (6.0%)
VTR + NC	5977	58 (1.0%)

Table 2. Statistics of sensitive words in the dictionary.

859 different actual sensitive words have been found among 13163 possible sensitive words examined. The numbers shown in the table are higher than these numbers because the composition of many words may fall into several categories. The proportion of sensitive words is surprisingly high. It is about 1% of all the items in our dictionary. In addition, a high proportion of sensitive words are very common words, as we can see in the examples in Table 3.

We can further observe the high percentage of sensitive words for the NC+VAUX

composition(38.4%). The main reason is that most words ending with 会 (association) and 能 (energy) may also be separated into two words, and once separated, 会 and 能 are usually interpreted as auxiliary verbs (会 - will, 能 - can).

Most sensitive words involve nouns (NC). The highest numbers of sensitive words are obtained from NC+VTR and ADV+NC compositions. In addition, the percentages of sensitive words of these compositions are relatively high.

From these observations, it becomes clear that we cannot consider the items in a dictionary as inseparable or equally separable. It is important to make distinction between sensitive and non-sensitive words. Precaution should also be paid when one adds new items into a machine readable dictionary for MT.

This study is not complete in the sense that several other possible compositions have not been examined. For example, we did not check the NC+NC composition. Intuitively, this is the composition the less affected by sensitive words, and many items in the dictionary (12571 of them) may correspond to this composition. However the sensitive problem does occur in this case, but in a different manner. The first NC may be syntactically associated with some precedent words before being associated with the second NC. For example, 便衣警察 (plain-clothes policemen) is an item in our dictionary. But this string should be separated into two words in the following sentence: 一旦穿上便衣警察就很难被认出 (Once in plain clothes, policemen can be hardly recognized). In this sentence, 便衣 (plain clothes) is the object complement of the verb 穿上 (wear), but 警察 (policemen) is the subject of another verb phrase 被认出 (be recognized).

5. Conclusions and Future work

In this paper, we described a new concept - sensitive word - in Chinese language. We first described how sensitive words may affect Chinese segmentation. The purpose of distinguishing sensitive words from non-sensitive words is to increase the efficiency of Chinese analysis and MT. We argue that if a segmentation solution fails to produce a translation, it is useless to try to break non-sensitive words into their components. On the other hand, sensitive words may be broken down, and this usually lead to a new syntactic structure for the whole sentence. Thus we have higher chance to reach at the correct interpretation of the sentence.

In order to see the scale of sensitive words in Chinese, we examined a machine readable Chinese dictionary. The number of sensitive words found is surprisingly high. We found more than 859 sensitive words among about 87 599 words and phrases. In other words, about 1% of dictionary items are sensitive words. This shows that sensitive words are widely spread in Chinese. It is worth more intensive study in the future.

We will continue this study by a more thorough analysis of the machine readable dictionary, and try to extract a set of common rules to detect sensitive words. Such a set of rules will be very useful when one establishes or enriches a Chinese dictionary.

Acknowledgment: This work has been partly supported by a grant for international Scientific Research (09044179) from the Education Ministry of Japan.

References

1. J.-S. Chang and e. al., Chinese word segmentation through constraint satisfaction and statistical optimization. ROCLING-IV, Taiwan, 147-165 (1991).
2. K.-J. Chen and S.-H. Kiu, Word identification for Mandarin Chinese sentences. 5th International Conference on Computational Linguistics, 101-107 (1992).
3. T.-H. Chiang and e. al., Statistical models for segmentation and unknown word resolution. 5th R.O.C. Computational Linguistics Conference, 123-146 (1992).
4. T. Dunning, Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, vol. 19, pp. 61-74 (1993).
5. W. Jin and J.-Y. Nie, Segmentation du Chinois - une Etape Cruciale vers la Traduction Automatique du Chinois. in La Traductique, P. Bouillon and A. Clas, Eds. Montreal: Les presses de l'Université de Montréal, pp. 349-363 (1993).
6. B.-I. Li and e. al., A maximal matching automatic Chinese word segmentation algorithm using corpus

tagging for ambiguity resolution. R.O.C. Computational Linguistics Conference, Taiwan, 135-146 (1991).
7. N. Y. Liang, The Automatic Segmentation in Written Chinese and an Automatic Segmentation System - CDWS. The Academic Journal of Beijing Institute of Aeronautics and Astronautics, vol. 4 (1984).
8. R. Sproat and C. Shih, A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, vol. 4, pp. 336-351 (1991).
9. R. Sproat, C. Shih, W. Gale, and N. Chang, A stochastic finite-state word-segmentation algorithm for Chinese. ACL'94(1994).
10. C.-L. Yeh and e. al, Rule-based word identification for Mandarin Chinese sentences - A unification approach. Computer processing of Chinese and Oriental Languages, vol. 5 (1991).
11. KaiYing L., Estimation Report of Chinese Word Segmentation, Chinese Computerworld, Vol.584, No.12, pp187-189(1996).

Composition	Example words	Example sentence
ADV + ADV	中共 (the Chinese Communist Party)	他/谈话/中/共/提到/三个/问题. (He mentioned 3 problems in his talk)
ADV + NC	后门 (back door)	我/进来/后/门/还/开着. (The door remained open after I came in)
ADV + PREP	反对 (oppose)	他/得到/职称/后/反/对/工作/不感兴趣/了. (He is no longer interested in the work once he obtained the title)
ADV + QUAN-CL	好几个 (several)	安顿/好/几个/人/后, ... (after settling down several people, ...)
ADV + VAUX	全能 (all-round)	他们/全/能/上/大学. (They can all go to university)
ADV + VINT	才干 (capability)	你/答应/了/这个/条件/他/才/干. (He accepts only if you agree with this condition)
ADV + VTR	中选 (be elected)	1 9 9 3 年/中/选/人民代表. (People's representatives are to be elected in the mid-1993)
CONJ + NC	同盟军 (federated army)	当地/民众/同/盟军/一同/作战 (The local civilians fought together with the federated army)
CONJ + VTR	并举 (be used together)	他/说到/一个/问题/, /并/举/了/桓 例子. (He talked about a problem, and gave an example of it)
NC + PREP	编者按 (editorial)	编者/按/作者/要求/作/了/修改 (The editor changed according to the author's requirement)
NC + VAUX	家长会 (association of parents)	学生/家长/会/不满意. (The parents of the students will be dissatisfied)
NC + VTR	机器翻译 (machine translation)	我们/什么/时候/才能/用/机器/翻译/文章 ? (When will we be able to translate papers by machine?)
PREP + NC	把门 (guard the door)	他/把/门/关上. (He closed the door)
PREP + VTR	从来 (ever)	从/来/这里/以后 ... (Since coming here, ...)
VPRED + PREP	想到 (think about)	我/想/到/那里/去. (I want to go there)
VPRED + VTR	不敢当 (not up to [a honor])	他/不敢/当/头. (He does not dare to be chief)
VTR + NC	管理学院 (Institute of management)	用/科学/方法/管理/学院 (manage the institute in a scientific way)

Table 3. Examples of sensitive words (the sensitive words should be broken in the examples).