Sensitive Words and their Application to Chinese Processing

Fuji Ren [1], Jian-Yun Nie [2]

[1] Faculty of Information Sciences, Hiroshima City University, Japan
ren@its.hiroshima-cu.ac.jp
[2] Department of Computer Science and Operation Research, University of Montreal, Canada
nie@iro.umontreal.ca

Abstract. Sensitive words are the compound words whose syntactic category is different from those of their components. According to the segmentation, a sensitive word may play different roles, leading to significantly different syntactic structures. If a syntactic analysis fails for a Chinese sentence, instead of examining each segmentation alternative in turn, sensitive words should be first examined in order to change the syntactic structure of the sentence. This will lead to a higher efficiency. Our examination of a machine-readable dictionary shows that there are a great number of such words. This shows that sensitive word is a widespread phenomenon in Chinese.

1 Introduction

Sensitive words refer to those Chinese compound words whose syntactic category is different from those of their components. For example, the string "*jiang lai*" may be segmented as a single word or as two words. In the first case, it is an adverb, meaning "in the future". In the second case, it becomes a sequence of an auxiliary verb and a verb, meaning "will come". The segmentation of sensitive words is critical in the determination of the syntactic structure of the whole sentence. However, this problem has been neglected in the past. People usually segment Chinese words using a longest-matching algorithm, which segment the above string systematically to a single word. Even if a backtracking strategy is used, the second segmentation will be suggested far later, after several other segmentation results have been examined. In this paper, we argue that sensitive words should not be dealt with in the same way as other segmentation ambiguities. They represent a possible breaking point that may change significantly the syntactic structure of the sentence. Therefore, if a segmentation result does not lead to a success in syntactic analysis, the sensitive words should be first checked. This may greatly improve the efficiency of Chinese analysis.

In this paper, we will first describe the problem of sensitive words in more detail, and their impact in Chinese segmentation. Then we will carry out a survey on a machine-readable dictionary. Our analysis shows that there are a great number of sensitive words in this dictionary. Some compositions are particularly productive in

sensitive words. This shows clearly that sensitive word is not an isolated, but widely spread phenomenon in Chinese. A correct processing is necessary.

2 Chinese segmentation and sensitive word

There have been two main groups of approaches to Chinese segmentation: dictionary-based approach [1, 4, 5, 8] and statistical approach [2, 3, 7]. We will only describe the first approach here. The second approach usually leads to similar results.

Dictionary-based approaches 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 that contains the longest (or equivalently, the fewest) words. For example, the sentence "*ji qi fan yi bu zhun que*" (Machine translation is not accurate) may be segmented, among others, in the following ways:

ji qi fan yi / bu / zhun que (machine translation, not, accurate = MT is not accurate)
ji qi / fan yi / bu / zhun que (machine, translate, not, accurate ˜ MT is not accurate)
ji qi fan yi / bu / zhun / que (machine translation, not, allow/accurate, indeed)

Among these possible solutions, the longest-matching algorithm will choose the first one. In most cases, the longest-matching method can choose the correct solution. Its accuracy is typically around 90% [6]. However, if the sentence is slightly changed to "*ji qi fan yi wen zhang bu zhun que*" (Using machines to translate articles is not accurate), then the algorithm will choose the following wrong segmentation:

ji qi fan yi / wen zhang / bu / zhun que (machine translation, article, not, accurate)

The correct segmentation should separate "*ji qi fan yi*" into two words. These examples illustrate the impact of segmentation to the whole Chinese analysis.

The problem of segmentation is its ambiguities. Many Chinese words containing at least two characters can be decomposed into simpler words (in modern Chinese). However, most such decompositions do not change radically the syntactic categories. Typically, a compound noun, if decomposed into Noun + Noun structure, would not change significantly the syntactic structure of the sentence. On the other hand, a decomposition may lead to two words with very different categories. For example, "*jiang lai*" (in the future – an adverb), if decomposed into "*jiang / lai*" (will, come – a auxiliary + verb sequence) would make a big difference in sentence's structure. It is this case that sensitive words are involved.

Among the compound words of the last case, we still distinguish two different cases:

1. Once decomposed, the words are unusual, or they cannot appear together;
2. The component words are usual words, and can follow each other in

normal Chinese sentences[1].

We only consider the words in the second case to be sensitive words. The string "*zhun* que" (accurate) is an example of the first case. "*zhun*" (accurate/permit) and "*que*" (indeed) separately are also words. However, they do not appear together in modern Chinese. This is not a sensitive word. On the other hand, it is easy to find cases where "*jiang lai*" (in the future) should be separated into "*jiang / lai*" (will, come). Therefore, it is a sensitive word.

Now, why sensitive words are interesting to study? Usually, Chinese processing follows the following pattern: one starts with the best segmentation result, if the subsequent analysis fails, we restart with the second best segmentation result, and so on. In this pattern, the quality of a segmentation result is solely based on word length. In fact, many segmentation ambiguities only affect word length, but not their meaning and the syntactic structure of the sentence. Taking such an alternative would not help at all the subsequent analysis if the first failed. Therefore, it is better to take an alternative that will lead to a significant change in sentence structure. This is the role of sensitive words: when a first analysis with the best segmentation result fails, we should first try to break down sensitive words involved. In this way, we can expect to arrive at the correct analysis result more quickly.

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 and it can be translated directly as a whole. Therefore, in many machine-readable dictionaries (in particular, for Chinese), long phrases are also included. Then how many words are sensitive words in such a dictionary? This is the question we will examine in the next section by analyzing an online Chinese dictionary.

3 Analysis of sensitive words in a machine-readable dictionary

The machine-readable Chinese dictionary we analyze here contains 87599 words (phrases). This is a dictionary used by a segmentation program, and manually checked by a Chinese linguist. The aim of the analysis is to show the extent to which sensitive words are spread in Chinese.

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

In many cases, compound words (phrases) are constructed from simpler words according to some rules. The following rules are commonly used:

NC + NC —> NC	(*gong si* + *zhi yuan* = employee of a company)
NC + V —> V	(*ren gong* + *fan zhi* = artificially breed)
NC + V —> NC	(*ji qi* + *zhe tu* = drawing by machine)
VTR + NC —> NC	(*jing* + *kou* = import)
VTR + NC —> VINT	(buo + *zhong* = weed)
V + V —> V	(*bian* + *yin* = edit and print)

[1] We use "normal sentences" because one can always imagine some very peculiar sentences for many word sequences. However, these sequences do not occur in normal texts.

Table 1. Symbols and their meaning

Symbol	Meaning	Example
ADJ	adjective	*nei* (inside)
ADV	adverb	*yi hou* (after)
CONJ	conjunction	*he* (and)
NC	common noun	*ji qi* (machine)
PREP	preposition	*bi* (in comparison with)
QUAN-CL	quantity + classifier	*ji ge* (several)
VAUX	auxiliary verb	*neng* (can)
VINT	intransitive verb	*xue xi* (study)
VTR	transitive verb	*guan li* (manage)
V	VINT or VTR	

Sensitive words are mainly due to the fact that a Chinese word (especially single-character word) may play different roles. For example, "*ba men*" (guard the door) corresponds to the rule V + NC —> VINT. However, "*ba*" is can also be a preposition that introduces the object in the structure "*ba* Something Verb", meaning "to Verb Something". The string "*ba men*" can also be considered as a sequence of "*ba* Something" as in "*ta ba men da kai*" (he opens the door = he *ba* door open). Therefore, "*a men*" is a sensitive word. In our examination of the MRD, 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, together with some examples that show normal situations where the word should be broken down.

Table 3 shows the number of sensitive words that we determined manually from a machine-readable dictionary for each word composition. The criteria used are those stated earlier, namely, 1) the component words may play significantly different roles from their compound in reasonable circumstances. 12 357 possible sensitive words have been identified according to the compositions from 87 600 words in the dictionary. Among them, 764 are actual sensitive words. The numbers shown in Table 3 are higher because some words fall into several categories. The proportion of sensitive words is surprisingly high. It is near 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 2.

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 *hui* (association) and *neng* (energy) may also be separated into two words, and once separated, ª· and ƒ‹ are usually interpreted as auxiliary verbs (*jiang* - will, *neng* - 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. It is important to make distinction between sensitive and non-sensitive words.

Table 2. Some compositions of sensitive words and examples

Composition	Example word	Example sentence
ADV + NC	*hou men* (back door // after, door)	ta / jin lai / hou / men / hai / kai / zhe I / enter / after / door / still / open = The door remained open after I came in
ADV + QUAN-CL	*hao ji ge* (several // well, several)	*an dun / hao / ji ge, ...* settle / well / several / person / after, = after settling down several people,
ADV + VAUX	*quan neng* (all-round // all, can)	*ta men / quan / neng / jing / da xue* they / all / can / enter / university = They can all go to university
ADV + VINT	*cai neng* (capability // then, do)	*ni / tong yi / le / ta / cai / neng / zuo* you / agree / already / this / condition / he / then / do = He accepts only if you agree with this condition
PREP + NC	*ba men* (guard the door // to, door)	*ta / ba / men / da kai* he / to / door / close = He closed the door
PREP + V	*cong lai* (ever // since, come)	*cong / lai / zhe li / yi hou …* from / come / here / after... = since coming here ...
VAUX + PREP	*xiang dao* (think about // think, to)	*wo / xiang / dao / na er / qu* I / think / to /there / go = I want to go there

Table 3. Statistics of sensitive words in the dictionary

Composition	nb. words with the pattern	nb. sensitive words (percentage)
ADV + ADV	401	10 (2.5%)
ADV + NC	2230	103 (4.6%)
ADV + PREP	294	9 (3.1%)
ADV + QUAN-CL	9	4 (44.4%)
ADV + VAUX	148	3 (2.0%)
ADV + VINT	1265	20 (1.6%)
ADV + VTR	1155	87 (7.5%)
CONJ + NC	131	14 (10.7%)
CONJ + VTR	94	24 (25.5%)
NC + PREP	628	45 (7.2%)
NC + VAUX	249	95 (38.2%)
NC + VTR	3994	252 (6.3%)
PREP + NC	748	39 (5.2%)
PREP + V	453	15 (3.3%)
VAUX + PREP	96	5 (5.2%)
VAUX + VTR	363	19 (5.2%)
VTR + NC	5976	57 (1.0%)

4. 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 leads 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 12,357 words which correspond to the composition patterns we determined among 87 600 words in the dictionary. Among them, 764 are actual sensitive words. In other words, about 0.9% 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.

This paper only concerns the phenomenon of sensitive word and its percentage in a MRD. We will continue our study by analyzing the proportion of sensitive words in Chinese texts. This will also give another useful indication on how each composition is productive in sensitive words.

References

1. Chang, J.-S., Chen, C.-D. and Chen, S.-D. Chinese word segmentation through constraint satisfaction and statistical optimization. ROCLING-IV, Taiwan, (1991) 147-165.
2. Chiang, T.-H., Chang, J.-S., Lin, M.-Y. and Su, K.-Y. Statistical models for segmentation and unknown word resolution. 5th R.O.C. Computational Linguistics Conference, (1992) 123-146.
3. Dunning, T., Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, vol. 19, (1993) 61-74.
4. Li, B.-Y., Lien, S., Sun, C.-F. and Sun, M.-S. A maximal matching automatic Chinese word segmentation algorithm using corpus tagging for ambiguity resolution. R.O.C. Computational Linguistics Conference, Taiwan, (1991) 135-146
5. Liang, N. Y. and Zhen, Y.-B. A Chinese word segmentation model and a Chinese word segmentation system PC-CWSS. COLIPS, 1(1) (1991) 51-55.
6. Liu, K.Y. Estimation report of Chinese word segmentation, Chinese Computer World, vol. 584, no. 12, (1996) 187-189
7. Sproat, R. and Shih, C. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 4(4) (1991) 336-351.
8. Yeh, C.-L. and Lee, H.-J. Rule-based word identification for Mandarin Chinese sentences - A unification approach. Computer processing of Chinese and Oriental Languages, 5(2) (1991) 97-118