Word Segmentation needs change

— From a linguist's view

Zhendong Dong
Research Center of Computer
& Language Engineering, CAS
dzd@keenage.com

Qiang Dong
Canada Keentime Inc.
dongqiang@keenage.com

Changling Hao
Canada Keentime Inc.
support@keenage.com

Abstract

The authors propose that we need some change for the current technology in Chinese word segmentation. We should have separate and different phases in the so-called segmentation. First of all, we need to limit segmentation only to the segmentation of Chinese characters instead of the so-called Chinese words. In character segmentation, we will extract all the information of each character. Then we start a phase called Chinese morphological processing (CMP). The first step of CMP is to do a combination of the separate characters and is then followed by post-segmentation processing, including all sorts of repetitive structures, Chinese-style abbreviations, recognition of pseudo-OOVs and their processing, etc. The most part of post-segmentation processing may have to be done by some rule-based sub-routines, thus we need change the current corpus-based methodology by merging with rule-based technique.

1 Introduction

Chinese word segmentation seems to be an old grandma's story. We very often hear some contradictory remarks about its advance. Most of reports from the evaluation tasks always gave us positive, or even impressive results, such as over 96% accuracy, but some reports were rather negative and expressed their deep concern. They claimed that word segmentation was still entangled in a difficult situation and no breakthrough in real applications. By careful and longtime observation, the incompetence is usually caused by the coarseness in the currently prevalent technology.

We carefully observed some Chinese-English MT systems and found some errors were caused even in the very early stage of the processing, that is, in the stage of word segmentation. No matter the MT is statistics-based or rule-based, they have their Achilles' heel in the segmentation stage. Can today's prevalent technology effectively cope with the problem? Or do we need some change? The present technology is characterized by its "trilogy", that is, "corpora + statistics (ML) + evaluation". We regret to say that many researchers today may be indulged in methodology itself rather than the language they have to target. They are enchanted by the scores and ranks, but they forget the object they are processing.

Therefore we propose that a Chinese morphological processing (CMP) should be taken to replace the current Chinese word segmentation. CMP includes the following components:

- Chinese character processing (CCP)
- Initial combination of Chinese multi-character expressions (CMEs)
- Morphological structure processing (MSP)

2 Chinese character processing

2.1 "Word" in Chinese

"Word or no word" may be an even older story in Chinese linguistic circle. One assertion about Chinese words may be quite popular, even to most of western researchers in the NLP circle, that is, different from English or other western languages, there is no space between Chinese words and thus segmentation of a running text into words is necessary for Chinese processing. However, do words really exist in Chinese? It is still a vexing and controversial issue. Some Chinese grammarians argue that in Chinese there are no words at all, but there are only characters instead and some express their strong objection.

What is a Chinese "word"? It was reported that the concept of "word" had not been introduced into China until the very beginning of the last century. In fact word is alien to Chinese. At least the concept of word in Chinese is rather vague. In Chinese there are no clear-cut distinction between characters and so-called word, either between multi-character words and those that are similar to English MWE. Ordinary English people may be surprised if they are told that even in popular Chinese dictionaries there are no entries equivalent to English "pork (猪肉)", "beef 牛肉)", "egg (鸡蛋)", "rain (verb 下雨)", "snow (verb 下雪)", but there are entries equivalent to English "lower limbs(下肢)", "give orders (下令)", "appendicitis (盲肠炎)". There is somewhat arbitrariness in recognition of Chinese "words", so the vocabulary in different Chinese dictionaries may vary very greatly. Does a dictionary take usage frequency into account when it decides on its entries? Let's compare their occurrence with the following entries in the dictionary as shown in Table 1. Let's compare the occurrence with the following entries in different dictionaries and in reference to Google's results. In Table 1, "-" indicates that the entry does not occur and "+" indicates the entry occurs.

Entries	3 Popular dictionaries	Results in Google
身为	- 现汉[1] - 规范[2] - 新时代汉英[3]	32,500,000
身亡	- 现汉 + 规范 - 新时代汉英	24,300,000
身居	- 现汉 + 规范 - 新时代汉英	16,600,000
身故	+ 现汉 - 规范 + 新时代汉英	6,760,000
身教	+ 现汉 + 规范 + 新时代汉英	497,000
身历	- 现汉 + 规范 + 新时代汉英	409,000
身受	+ 现汉 + 规范 + 新时代汉英	900,000

Table 1. Comparison of entry occurrence in dictionaries

In a word, since "word" in Chinese is rather vague, what is a better tactics we should take then? The present word segmentation is burdened too heavily. In comparison with English tokenization, it goes too far. Does English tokenization deal with MWEs, such as "United nations", "free of charge", "first lady"? Why does Chinese word segmentation have to deal with Chinese multi-character "word"?

2.2 Chinese character processing (CCP)

We propose that the real task of so-called Chinese word segmentation is to segment a running text into single characters with spaces between. We call this processing Chinese character processing (CCP). CCP is in parallel with English tokenization. In most cases CCP can achieve 100% accuracy. The most important task for CCP is not only to segment a text, but also to obtain various kinds of information (syntactic, semantic) of every character. What will be followed depends on the tasks to be designated. Usually a demand-led morphological processing will be taken.

3 Initial combination

In most cases, what we called initial combination of Chinese multi-character expressions (CMEs) should be followed indispensably. It may be either shallow or deep, and may be done either with the help of a lexical database or a corpus, and the longest matching may be the frequently-used technique.

[1] Modern Chinese Dictionary
[2] Modern Chinese Standard Dictionary
[3] New Age Chinese-English Dictionary

4 Morphological structure processing (MSP)

4.1 Pseudo-OOVs

The first task of MSP is to recognize and process Chinese OOVs. What are OOVs in English? Normally if a string between two spaces in a running text does not exist in the lexical database or the corpus the processing system is using, this string is taken as an OOV. However, what is an OOV in Chinese then? It is really not so easy to define an OOV in Chinese as in English. The recognition of English OOVs may be done in the phase of tokenization, but the recognition of Chinese OOVs should, in a strict sense, not be done in so-called word segmentation. It should be regarded as a special phase of the morphological processing. It is commonly acknowledged that OOV recognition is the most serious factor that impairs the performance of current Chinese word segmentation.

We may first look at some instances of machine translation results and find the actual problems. The reason why we use MT systems to test and evaluate segmentation is because this will make it explicit and easy for human to assess. One error in segmentation makes a 100% failure in translation. In our examples, the translation (a) is done by a statistical MT system and the translation (b) by a rule-based MT system. (C) is human translation, which may help make comparison and find the errors made by MT.

1. 美国民众**力挺**南京申办 2020 年奥运会。

(a) Americans even behind the bid to host the 2020 Olympic Games in Nanjing.

(b) American people's strength holds out in Nanjing and bids for the 2020 Olympic Games.

(c) Americans fully backed up Nanjing’s bid to host the 2020 Olympic Games.

Chinese OOVs can be roughly categorized into two classes, one is true OOVs and the other is pseudo-OOVs. The recognition and processing of true OOVs can be done as English OOVs are treated in English. However, the recognition and processing of Chinese pseudo-OOVs should be done by a special processing module. Chinese pseudo-OOVs includes two types: plain pseudo-OOVs, such as “力挺”, “洁肤”, “野泳”, “浴宫”, “首胜”, “完胜”, and abbreviated pseudo-OOVs, such as “二炮”, “世博”, “严打”, “婚介”, “疾控中心”, “驻京办”, “维稳办”, “园博会”, “中老年”, “事病假”, “军地两用”.

- **Plain pseudo-OOVs**

A pseudo-OOV is a combinatory string of Chinese characters in which each character carries one of its original meanings and the way of combination conforms to Chinese grammatical pattern. In the above Chinese sentence the word “力挺” is a typical pseudo-OOV. “力挺” is a combination of two characters, “力” and “挺”. “力” has four meanings, one of which is “do one’s best”. “挺” has six meanings, one of which is “back up”. Originally in Chinese dictionaries we can find the following expressions similar to the pattern of “力挺”, such as “力避”, “力持”, “力促”, “力挫”, “力荐”, “力戒”, “力克”, “力拼”, “力求”, “力图”, “力争”, “力主”. In all these expressions the character “力” carries the same meaning as that in “力挺”, and the second characters in the combinations are all actions. Therefore the expression “力挺” is a grammatical and meaningful pseudo-OOV. It should be noticed that this kind of pseudo-OOV is highly productive in Chinese. In addition to all the dictionary entries that we listed above, we found “力陈(to strongly state)”and “力抗(to strongly resist)” are already used in the web. Its highly occurrence in real texts calls our special attention. Let’s see how MT will tackle them poorly.

2. 辩护人**力陈**多处疑点。

(a) Chen multiple defense of human doubt.

(b) Many old doubtful points of the manpower of pleading.

(c) The pleader argued and showed many doubtful points.

We wonder how the current technique of segmentation tackles the problem. We are not sure how one error in a segmentation effect the score in Bakeoff.

Let’s look at two more examples and have a brief discussion of them.

3.据邻居反映，案发当天中午有一个快餐**外卖郎**来过被害人家中。

(a) According to neighbors reflected the incident that day at noon there is a fast food take-Lang came to the victim's home.

(b) According to the information of neighbour's, a fast food takes out the my darling to been to victim's home at noon on the day when the case happened.

(c) According to the neighbors, at noon on the same day a fast food takeout boy came to the victim’s house.

4. 一个官员被**修脚女**刺死了。

(a) One officer was stabbed to death the women pedicure.

(b) An officer is trimmed the foot daughter and assassinated.

(c) An official was stabbed to death by the girl pedicurist.

All the four erroneous MT translations above originate from the so-called recognition of OOVs “外卖郎” and “修脚女” in the segmentation. The MT systems might make out “外卖”and “郎” or “修脚” and “女” separately, but fail to recognize their combinations. The combination pattern of these two plain pseudo-OOVs is a very typical and popular one in Chinese, just similar to the suffix “-er” or “-or” in English to derive a noun of a doer. “外卖郎” is a combination of “外卖”(takeout) and “郎”(boy). When a MT failed to tackle it, the translation would be so poor.

- **Abbreviated pseudo-OOVs**

Different from English abbreviations or acronyms, Chinese abbreviations in essence are contracted forms of words and expressions. The contraction is mainly related to three factors: (1) maximal preservation of the original meaning; (2) possible maintenance of Chinese grammatical structural pattern; (3) consideration of acceptableness of rhythm. Let’s take “维稳办” for example. “维稳办” is the contraction of “维护稳定办公室”. The literal translation of the expression is “maintain stability office”. Thus the first part of the expression “维护稳定” is contracted to “维稳”, and the second part is contracted to “办”. “维护稳定” grammatically is a “verb + object” structure while “维稳” can be regarded as the same grammatical structure. Grammatically “办公室” is modified by “维护稳定”, and in the contraction the word “办” is also modified by the contraction “维稳”. As for acceptableness of rhythm, “维稳办” is a three-character expression, in which the first two are a “verb + object structure and the last is single. The structure of “2-character verb + 1-character noun” is a highly-productive pattern of noun expression in Chinese. So it is desirable to process this type of structures before syntactic processing. As the structure can usually be patternized, it is possible to have them well-processed. We propose that we should deal with it in the morphological processing stage.

4.2 Repetitive structures

First let’s look at a MT translation and see what has happened when a Chinese repetitive structure is ill-processed.

5. 你来**穿穿看**，太小了。

(a) Come see Chuan Chuan, too small.

(b) You come to wear looking, it is too small.

(c) Come and try on, it is too small.

The above two erroneous MT translations (a) and (b) originate from the failure in dealing with a typical verb structural pattern for expression to urge someone to have a try. This pattern is: “VV看”, its actual meaning is “have a try” and “to see if …”. The literal translation of the above instance “穿穿看” may be “put on, put on and let’s have a look”. Similarly we can have “吃吃看” (which can be literally translated as “taste, taste, and let’s see”).

Chinese is unique with its various types of repetitive structures. They are by no means rare phenomena in real texts. Any negligence or failure in the processing of repetitive structures will surely spoil the succedent tasks. Unfortunately this problem has not caught enough attention of researchers and developers of word segmentation tools. Most of neglecters usually leave the problem to the vocabulary that they collect. Let’s compare the following two groups of translations:

Group A

你再仔细听一听，是不是哪里漏水了。

他看了看停在旁边的火车。

Group B

你再仔细嚼一嚼，是不是有薄荷味。

他坐了下来，又向后靠了靠。

Group A1

You listen carefully, is not where the leak was.

He looked at the stop next to the train.

Group B1

Carefully you chew a chewing is not a mint flavor.

He sat down, then back by the by.

The English translations of the repetitive structures in Group A1 are acceptable for the structures “听一听” and “看了看” are no doubt in the vocabulary. And the translations of Group B are messy enough to show that the repetitive structures become OOVs and are not well-processed.

Generally most of Chinese repetitive structures originate from three word classes:

- Verb repetitive patterns:

AA	听听, 想想, 谈谈
ABAB	商量商量, 研究研究
A一/了A	嚼一嚼, 看了看
AA看	穿穿看, 吃吃看
A了一/又A	闻了一闻, 按了一按,摸了又摸

- Adjective repetitive patterns:

AA	大大, 轻轻, 红红, 胖胖
AABB	漂漂亮亮, 大大方方,斯斯文文
ABAB	白胖白胖, 焦黄焦黄

- Classifier repetitive patterns:

AA	个个（是好汉）, 件件（是稀世珍宝）
一AA	一辆辆, 一只只, 一碗碗, 一床床
一A一A	一件一件, 一套一套,一块一块
一A又一A	一张又一张, 一朵又一朵, 一条又一条

All these patterns are highly productive in Chinese. It will be impracticable for any Chinese parsing or MT systems to leave all the resolutions of them to the vocabulary rather than special processing module.

4.3 Plain classifier and unit structures

Chinese is featured by its plenty of classifiers. In many cases a concrete noun occurs idiomatically with its particular classifier especially when modified a numeral, for example, “一个人”(a person), “两辆车”(two cars), “三公斤苹果”(3 kilos of apples). The processing of this type of structures will surely benefit the succeeding parsing and even word sense disambiguation. Besides the processing is comparatively easy even in the early stage.

4.4 Chinese verb aspect processing

The verb aspect in Chinese is different from that in English. In general, by using Chinese aspects, we add some procedural tune to a verb rather than relating to time. In other words Chinese verb aspects give hints of the developmental phases or results, or the capability or possibility of the events. Chinese verb aspects are expressed by the aspect markers, such as simple markers “上”, “下”, “进”, “出”, “回”, “过”, “起”, “开”, “到” and compound markers “上来”, “下去”, etc.

Again let’s look at two pair of Chinese-to-English MT translations.

(6) 要干的工作太多了，一个人实在是干不过来了。

(a) To dry too much work, a person indeed dry However come.

(b) The ones that should do have too much work, one can not really be dry.

(c) I have too much work to do, I can hardly cope with it.

(7) 姑娘说着说着哭起来了。

(a) Said the girl spoke to cry.

(b) The girl has cried saying.

(c) The girl began to weep while talking.

The messy translations tell us how serious the impairment of the translation will be if we fail to process the Chinese verb aspects.

Table 2 shows the meanings conveyed by most Chinese aspect and its corresponding "aspect markers" and examples. Finally, when speaking about Chinese aspect, one point we would like to invite readers' attention that different from the aspect of English. It is known that English aspect is usually closely related to tenses, for example, English verbs can be used in progressive aspect with various tenses, such as present progressive, progressive and future progressive tenses. However, Chinese aspects are related to the development of the event itself, but not related to the time when the event happens.

5 Conclusion

Is it time for Chinese NLP circle to rethink what we have actually achieved in the word segmentation and consider some radical change? How much room left is there for the current trilogy to improve? We propose that we should have morphological processing to replace the so-called word segmentation. We have designated new tasks for the processing. In addition, we hope that we should design and use a new evaluation method. The general idea of new evaluation is to use a post-segmentation, or post-morphological-processing task, say, chunking, to evaluate, rather than the present method of isochronous self-testing.

sememe in HowNet	meaning	marker	examples
{Vsuppose|假定}	presupposing	起来	读~流畅
{Vstart|发端}	inceptive	起来	双方对骂~
		上	在一旁聊~了
{Vgoingon|进展}	progressive	在	~发言呢
		正	~睡觉呢
		正在	~干活
		着	说~说~动手了
{Vcontinue|延续}	protractive	下去	谈~会有结果
{Vend|完结}	terminative	过	吃~饭再走吧
{Vachieve|达成}	perfective	出	做~新成绩
		出来	算~了吗
		到	接~人了吗
		得	饭做~了
		过来	错的地方改~
		过去	被我蒙~了
		好	功课做~了
		见	听~了但看不~
		上	吃~一顿饱饭
		下	谈~那笔生意
		着	见~要见的人
{Vable|能力}	capable	得到	办~
		得过	信~
		得过来	忙~
		得了	一个人干~
		得起	买~
		得下	装~
		起	输~输不~
		下	可以睡~3 个人
{Vincapable|没能力}	incapable	不得	动也动~
		不过	说~你
		不过来	一个人忙~
		不了	一个人可干~
		不起	负担~
		不下	吃~
{Vpossible|可能}	possible	得	这菜吃~吃不~
{Vtry|试试}	Trying	看	穿穿~

Table 2. Chinese aspect markers and their meanings

References

Hai Zhao and Chunyu Kit, 2008. Unsupervised Segmentation Helps Supervised Learning of Chinese Tagging for Word Segmentation and Named Entity Recognition. In Prceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008, Hyderbad, India.

Hwee Tou Ng and Jin Kiat Low, 2004. Chinese Part-of-speech Tagging: One-at-a-Time or All-at-once? Word-Based or Character-Based? In Proceedings EMNLP.

Nianwen Xue, 2003. Chinese Word Segmentation as Character Tagging. International Journal of Computational Lnguistics and Chinese Language Processing, 8(1):29-48

Wenbin Jiang and Haitao Mi and Liang Huang and Qun Liu, 2008b. Wird Lattice Reranking for Chinese Word Segmentation and Part-of-speech Tagging. In Proceedings of COLING

Xinnian Mao, Yuan Dong and Saike He, Sencheng Bao and Haila Wang, Chinese Word Segmentation and Name Entity Recognition Based on Condition Random Fields, In Prceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008, Hyderbad, India.

Zhendong Dong and Qiang Dong, 2006. HowNet and the Computation of Meaning, World Scientific Publishing Co. Pte. Ltd., Singapore

黄昌宁, 赵海, 2007, 中文分词十年回顾. 中文信息学报, 2007, 21(3):8-20.

黄居仁, 2009, 瓶颈, 挑战, 与转机: 中文分词研究的新思维. In Proceedings of CNCCL-2009, Yantai