Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 212-217

Chinese Word Segmentation

*Li Haizhou and Yuan Baosheng**

Kent Ridge Digital Labs, Singapore

Chinese word segmentation has been a very important research topic not only because it is usually the very first step for Chinese text processing, but also because its high accuracy is a prerequisite for a high performance Chinese text processing such as Chinese input, speech recognition, machine translation and language understanding, etc. This paper gives a review on the development of Chinese word segmentation techniques that have been applied to various applications on Chinese text processing. As the methodology varies in a very wide range according to its applications, in this paper it is viewed in terms of the knowledge resources on which segmentation methods based. We summarize the methods into two categories, that is, lexical knowledge based and linguistic knowledge based methods.

1. INTRODUCTION

It's well known that Chinese is an ideographic language and there is no word delimiter between words in written Chinese sentences. Therefore word segmentation becomes the very first task when processing Chinese text and in turn the accuracy of word segmentation is essential to the performance of the following procedures. As such, it has been widely studied in recent years and there have been quite a few publications on it[2-6,8-10,12-14,19,21,25]. As the definition of Chinese word itself is a complicated issue[1,11] and worth a separate paper, here we only focus on word segmentation techniques themselves trying to highlight the achievements so far obtained and the problems to be solved in the future. It's not easy to give a concise summary based on applications since we note that most of the methods are application oriented and driven by their respective needs. In this paper, we attempt to reveal the underlying technologies and give a comparison among them.

Of the algorithms or methods reported on Chinese word segmentation[2-7,8-10], most of them roughly fall into two categories: lexical knowledge based methods and linguistic knowledge based methods. Since one important task for Chinese word segmentation is to reduce unexpected segmentation errors, it' s worthwhile to discuss word segmentation errors first before going to details of the algorithms.

There are two major segmentation problems that might affect the accuracy of word segmentation performance, unknown word and word ambiguity. The first problem always happens as long as the text corpus contains new words. As a class of unknown word or new words usually shares similar attributes, new words could be identified by using heuristic rules, called new word detection. The second problem was found to consist of two typical types of ambiguities[2], i.e. words overlapping ambiguity and composition ambiguity.

As we know, the new word problem always happens no matter how big a lexicon is in a real application. The two major causes of segmentation errors usually come together leading to unexpected segmentation results. Typical examples of errors are insertion, deletion and misplacement of segmentation boundaries: (a) an insertion between 人民and 币, which indicates a new word problem; (b) a deletion between 矛 and 盾 in 他 卖 矛盾, which results in a composition segmentation error; (c) two misplacements in 确 实在 理 instead of 确实 在理, which leads to overlapping ambiguity segmentation errors. To discover these ambiguities and disambiguate them is a very difficult but yet very important task[6,10,26].

Lexical knowledge based segmenter only makes use of the lexicon knowledge to conduct its segmentation judgment. It is efficient and straightforward in practice. Resorting to more powerful decoding approaches, linguistic knowledge based segmenters combine lexical information with language statistical information to provide better accuracy. The

* Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613, hzli@krdl.org.sg, baosheng@krdl.org.sg

lexical knowledge based approach effectively deals with all the ambiguity problems by ignoring it. Linguistic knowledge based approach alleviates the problems by introducing contextual constraints into the segmenter.

In the next section, we will give comprehensive review on various algorithms that use a single lexicon as the sole resource to conduct the segmentation. The section 3 summarizes segmentation methods that apply more linguistic knowledge. Finally, we summarize the discussions.

2. LEXICAL KNOWLEDGE BASED METHOD

Most of the earlier work on Chinese word segmentation is lexical knowledge based methods[2,8,10,12] which use lexicon as the only resource to conduct the segmentation. Thanks to its simplicity and efficiency, a considerable success has been reported.

2.1. Maximum matching (MM)

With maximum matching, a character string is compared with the entries of a lexicon so that all the substrings constituting lexicon items are highlighted. The principle of maximum matching, or called longest match, is to find the best segmentation with fewest and longest words among all the possible substring chains. The algorithm starts from the beginning of a sentence, finding the longest matching word and then repeating the process until it reaches the end of the sentence. It is noted that as long as the lexicon is a super set of a minimal complete set, for example, a word set comprising all the single character words, the algorithm will lead to an unique solution[21]. An unique solution does not mean a good solution. Unfortunately, as the segmentation is determined locally, the resulting sentence segmentation is always a suboptimum.

The MM approach always leads to one segmentation pattern resolving the ambiguity problem by ignoring it. As is the nature of the method, MM by itself is unable to deal with the composition segmentation errors. For example, segmenting sentence 这些学生会游泳, MM always gives an incorrect segmentation 这些 学生会 游泳 as a result of the fact that it favors longer word 学生会 over short words 学生 and 会. It does not reveal any overlapping ambiguity either, one can see that the ambiguity is still a big problem in practice. Using an application specific lexicon, it is possible to get rid of some ambiguities. Maximum matching with backtracking (MMB) is one of the variants of MM approaches to dealing with the ambiguity problems. MMB determines a segmentation according to a current word matching and its word matching history on its way[21]. Other approaches make use of linguistic knowledge and language statistics therefore they are able to deal with the ambiguity problems more effectively, which will be discussed in the next section.

Although both the rule and the implementation of the MM algorithm are very simple, this method gives a reasonable performance provided the lexicon used is appropriate for the task. Some previous studies shown that the size of a lexicon is even less important than the appropriateness of the lexicon to the particular corpus[20].

2.2 Forward and backward MM method

It is noted that MM method could begin from end of a sentence, conducting the segmentation in a backward pass. it might lead to a distinct segmentation result from that by forward MM. Therefore, forward and backward MM method, combining both forward MM and backward MM, is an alternative to discover the segmentation ambiguity.

This method carries out the segmentation in two steps: 1) getting segmentation results, or word chains, with both forward and backward MM; 2) resolving results from the two word chains according to some criteria. By comparing the two word chains, what is usually adopted is to keep the common segments in the two chains, then apply some heuristic rules or language knowledge to the conflict segments to single out the results. It was shown that FBMM is able to reveal most of the word overlapping ambiguities happening in both MM passes. It should be noted that FBMM is unable to indicate all the word overlapping ambiguities and composition ambiguities remained unresolved[21].

It is also noted that FBMM doubles the computation of MM method. Considering a word matching as a search process, one could speed up the MM procedure by sorting the lexicon in descending order by word frequency. This fast matching approach is applicable to all MM procedures. It's a good approach for those tasks where speed is a concern in particular.

2.3 Exhaustive words matching

This method, EWM, tries to get all the possible word boundaries for a sentence being processed and use dynamic search procedure to select one segmentation path with minimum number of segmentation notes from the segmentation lattice[10]. As multiple possibilities are available, one has to provide a criterion for the segmenter to come out with a choice. For example, least segments, maximum word length etc. This method seems promising but in fact it does not necessarily outperform other methods, say MM, because EWM does not use any additional linguistic knowledge besides the lexicon entries. As a result, all three types of segmentation errors still occur in the process. In addition, it also introduces much more computation and requires more memory for the dynamic searching. However, this method provides all the possible word boundaries and therefore could be good method when additional linguistic knowledge are available and a task requires doing so.

2.4 Using heuristic lexicon rules

There are some approaches which resort to constraint satisfaction based on syntactic or semantic features[22] and lexical heuristics [3-5,23,24]. One example is to use word boundary hints. In addition to punctuation marks, there are some other indications that could be helpful to improve segmentation accuracy. Those indications are heuristic linguistic knowledge such as those characters 1) that can be used only to head a word; 2) that can be used to end a word and 3) that can only be used as a single character word.

Some kind of part of speech taggings are also considered as lexical heuristics. In [23], a lexicon consists of words that are tagged to be word prefixes, word bodies and word suffixes. A productive structure to form new word is introduced in the way of prefix-body-suffix subject to the part of speech constraints.

As linguistic knowledge provides more precise information of word relationship in a text context, there is sizable literature on linguistic knowledge based methods for Chinese word segmentation[3-5,24,25]. It is shown that for application specific task such as new word detection for proper names, foreign names and place names, lexical rules play an important role in a high performance segmentation system[3-5]. Next we will summarize some of the work under Viterbi framework which is considered as a mainstream approach.

3. LINGUISTIC KNOWLEDGE BASED METHOD

Linguistic knowledge based approaches still rely very much on the lexicon. They usually start with all possible segmentations of a sentence, then pick the most likely segmentation from the set of possible segmentations using a probabilistic or cost-based scoring mechanism. The simplest approach, for instance, scores the paths based on the word frequency and picks the sentence with lowest cost as is described by Chang, Chen and Chen[13]. Approaches differ by their scoring or path searching processes. In addition to the word frequency, that is word unigram, some other information is also used to rank the possibilities. The literature involves part-of-speech information, morphological analysis, Chinese proper name automata[14], etc.

3.1 Using word unigram

In [14], Chinese word segmentation is viewed as a stochastic transduction problem. A dictionary or lexicon is represented as a Weighted Finite State Transducer[14], or WFST. The weight associated with a word is its word unigram. The summed unigram cost over all the possible paths are evaluated and the path with lowest cost is selected as the output sequence. The decoding process is a typical instance of Viterbi algorithm. As only the word unigram is used, the segmenter under discussion here is a zeroth-order model.

There are a number of words that are not caught by the lexicon, such as words that make up dates, numbers, proper names, place names, morphological structures, etc. One solution to the morphological problems is to build productive morphological processes into a WFST by introducing transition weights between the bodies and the affixes, such as nouns and their plural noun formation suffix 们. As such, WFST augments the dictionary to accommodate as many as possible the morphological decompositions. A WFST is also proposed to detect Chinese proper names by Chang et.[16] in a statistical manner. The approach does not preclude the extensions to other classes of words, such as transliterations of foreign words and Chinese place names. The additional WFST mechanisms could be viewed as an extension of the baseline dictionary lexicon into a orthographic lexicon.

3.2 Using word bigram and trigram

Usually, the segmentation ambiguity can not be resolved locally. Resorting to N-gram, one introduces more contextual constraints which will help a segmenter make a decision based on broader context. Following is a typical example:

1.(a) 他 说 的确 实在 理

(b) 他 说 的 确实 在理

The segmentation in 1(b) give a better total cost than 1(a) by the context. It is known that bigram and trigram are more practical to serve as the high order language models than the other N-grams. Using unigram alone, the segmenter might come out with 1(a) as the result because the accumulated unigram cost for 1(a) is lower. Applying unigram and bigram, 1(b) wins the scoring.

As indicated by Liang[2], there are two cases of unexpected segmentation. One is overlapping ambiguity where a character could go either way to form two words, such as 实 in example 1. Another is composition ambiguity where the subsegmentation is possible:

2.(a) 这些 学生会 游泳

(b) 这些 学生 会 游泳

One can find that 学生会, 学生 and 会 all are possible word entries, thus both results are valid based on lexicon entries. Without using language models, the segmenter will pick out incorrect segementation (a) as its output result because it has fewer words in the sentence satisfying the principle of MM criterion. However, a segmenter would get the correct segmentation (b) as its result if a bigram or trigram is used. It shows that higher order language models help remove the unexpected segmentations and therefore further improve segmentation performance.

3.3 Building a word lattice from a character string

Given a lexicon, one can easily construct a word lattice from a character string where all the possible word segmentation results are retained. Each word is associated with a unigram. Similarly, each word transition is associated with a word or word class* bigram[17]. Viterbi algorithm is implemented to decode the best path with least cost, which takes the word unigram and bigram into accounts. In some applications where N-best paths are required, Stack decoder is an alternative algorithm. In [17], the word lattice is passed to stack decoder to come out with an N-best list by using unigram and bigram. Thereafter, word class trigram is used to rescore the N-best list. The reason why the trigram is not incorporated into the stack decoder search at the first stage is simple: for the sake of searching space and decoding time.

It is presumed that any character could serve as a word boundary, therefore, the word lattice is built in a character-synchronized way. The number of lattice entries are pruned down on the way of decoding to get rid of unreasonable partial paths. The pruning reduces the search space and consequently speeds up the process.

3.4 The Viterbi framework

It is known that Viterbi algorithm conducts an optimal decoding through a word lattice. One can easily find that quite a few word segmentation approaches are basically derived from Viterbi framework and are given different names.

* A class of words consists of a group of words sharing common part of speech attributes.

Maximum matching is an extreme case of Viterbi considering that it only keeps one extension path when traversing forward or backward. Exhaustive matching comprises several derivatives of Viterbi procedures under various searching criteria: Minimum segmentations is a Viterbi procedure under least word transition criterion; Maximum word length is under maximum average length rule. Word unigram and part-of-speech tag information are also used in some MM applications to serve as additional knowledge to the scoring[14,15]. Word statistics are helpful because either least word transition or maximum average length alone by itself usually results in candidates of equal costs. It is noted that only the best choice is to be singled out in most of applications.

Within the Viterbi framework, it is also easy to apply various constraints and to achieve a balance among them. For example, high word transition cost tends to come out with less segments; highly weighted word unigram favors the most frequent words, etc. A careful weighting among the costs will direct the segmentation judgments.

3.5 Multi-stage word segmentation

It is shown[19] that a hybrid approach, combining statistics and lexicon knowledge, is most promising in terms of resolving the unknown word problems. Some of the hybrid methods involve multi-stage process to apply linguistic knowledge once at a time. The linguistic knowledge could be at different resolutions or for different purposes. As in [17], lexicon, unigram and bigram are used in the first pass to decode N-best segmentations, trigram is then used to reorder the resulting N-best segmentations. Another example is given in [20] where a generic word segmenter is followed by an unknown word detection mechanism. After the first step the segmented text is further analyzed by a set of heuristic morphological rules. A number of hybrid methods tailored to specific applications have been reported with success.

4. SUMMARY

As there is no unique solution to all the segmentation problems, Chinese word segmentation systems are usually application specific. In some applications where the processing time is a critical issue, one might sacrifice some accuracy. For example, to serve as the front-end of a N-gram statistics generator, the lexical knowledge based method is likely to be adopted while ignoring new word problems. When segmentation accuracy is a particular concern, as in Chinese spell check systems and Pinyin to Hanzi conversion systems, mechanisms with higher complexity are introduced to deal with the ambiguity problems[17]. More precise linguistic knowledge with appropriate decoding mechanism will ensure a better result. The Viterbi framework is a typical example to demonstrate how linguistic knowledge of different kinds at different levels contribute to Chinese word segmentation process. It is noted that the statistical modeling allows us to train up a language model instead of to program it while heuristic rules provide solutions to particular problems effectively. Therefore, the hybrid system is always an alternative in different tasks. To conclude, lexicon building, heuristic rules and linguistic statistics are still the key issues in the future study of Chinese word segmentation.

REFERENCES

1. 刘源、谭强沈、 旭昆(1994), 《信息处理用现代汉语分词规范 及自动分词方法》, 清华大学出版社 1994年6月第一版. pp1~10
2. 梁南元(1987), 书面汉语自动分词系统-CDWS, 《中文信信学报》, 1987年 第2期, pp45-47.
3. 孙茂松、张维杰(1993), 英语姓名译名的自动识别, 《计算语言学研究与应用》, 陈力为 主编, 北京语言学院出版社, 1993年10月 第1 版. pp144~49
4. 郑家恒、刘开瑛 (1993), 自动分词 系统中姓氏人名处理策略探讨, 《 计算语言学研究与应用》 陈力为主编, 北京语言学院出版社, 1993年10月第1版.
5. 宋柔等(1993), 基于语料库和规则库的人名识别方法, 《 计算语言学研究与 应用 》 陈力为 主编, 北京语言学院出版社, 1993年10月第1版. pp 150-54.

6. 白栓虎(1995), 汉语词切分及词性自动标注一体化方法, 《计算语言学进展与应用》 陈力为 袁琦 主编, 清华大学出版社, 1995年10月第1版. pp56-61.
7. Feng Zhiwei(1995), Seminar on Chinese Linguistics, DISCS, National University of Sinapore,1995. chp 3.
8. 梁南元(1987), "再论汉语自动分词和切词知识", 《中文信息处理国际会议论文集》, 1987年8月, 北京.
9. 何克抗、徐辉、孙波(1991), 书面汉语自动分词专家系统设计原理, 《中文信息学报》, 第5卷1991年第2期
10. 王小龙等(1989), 最少分词问题及其解法, 《科学通报》, 1989 年 第13期.pp1030-32
11. 刘源(1992), 字词频统计与汉语分词规范《语文建设》, 1992年, 第2期 pp. 35-38.
12. 揭春雨、刘源、梁南元, (1989) "论汉雨语自动分词方法", 《中文信信学报》, 1989 年 第3期.
13. Chang Jyun-Shen, C.-D. Chen and Shun-De Chen "Chinese Word Segmentation through constraint satisfaction and statistical optimization", Proc. of ROCLING IV, ROCLING, Taipei, pp 147-165
14. Richard Sproat, Chilin Shih, William Gale and Nancy Chang (1996), "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese", Computational Linguistics, Vol 22, Number 3, 1996
15. Jian-Cheng Dai and Hsi-Jian Lee, "Paring with Tag Information in a probabilistic generalized LR parser" (1994), International Conference on Chinese Computing, Singapore, pp33-39
16. Chang, Jyun-Shen, Shun-De Chen, Ying Zhen, Xian-Zhong Liu and Shu-Jin Ke (1992), "Large-corpus-based methods for Chinese personal name recognition", Journal of Chinese Information Processing, 6(3):7-15
17. Li Haizhou et al (1997), "Pinyin Streamer: Chinese pinyin to hanzi translater", Apple-ISS technical report
18. L. Rabiner and B.H. Juang (1993), "Fundamentals of speech recognition", Prentice Hall
19. Jian-Yun Nie, Wanying Jin and Marie-Louise Hannan, "A hybrid approach to unknown word detection and segmentation of Chinese" (1994), International Conference on Chinese Computing, Singapore, pp326-335
20. Fung Pascale and Wu Dekai (1994), "Statistical augmentation of a Chinese machine readable dictionary", WVLC-94, Second Annual Wordshop on Very Large Corpara
21. 马晏(1996), 基于评价的汉语自动分词系统的研究与实现, 《语言信息处理专论》黄 昌宁 夏莹 主编 清华大学出版社, 广西科学技术出版社, 1996年4月 第1 版. pp2-36. 1996.
22. Yeh Ching-long and His-Jian Lee (1991), "Rule based word identification for Mandarin Chinese sentences – a unification approach", Computer Processing of Chinese and Oriental Languages, 5(2):97-118
23. Baosheng Yuan, Yuqing Gao et al.(1996), "Chinese Dictation Kit: A Very Large Vocabulary Mandarin Speech Input System", ICCC'96, pp1-4. 1996.
24. Wang Yongheng, Haiju Su and Yan Mo (1990), "Automatic processing of Chinese words", Journal of Chinese information Processing 4(4):1-11
25. Wu Zimin and Gwyneth Tseng (1993), "Chinese text segmentation for text retrieval: Achievements and problems", Journal of the American Society for information Science, 44(9):532-542
26. 侯敏、孙建军、陈肇雄(1995), 汉语自动分词中的歧义问题, 《计算语言学进展与应用》 陈力为 袁琦 主编, 清华大学出版社, 1995年10月第1版. pp81-87.