Word Boundary Information and Chinese Word Segmentation[1]

Gaoqi Rao Endong Xun
International R & D Center for Chinese Education
Beijing Language and Culture University, 15 Xueyuan Road, Haidian District,
Beijing, P. R. China, 100083,
{raogaoqi, edxun}@blcu.edu.cn

Abstract

Chinese word segmentation could be considered as a problem of word boundary recognition. Word boundary information plays a significant role in human language acquisition and automatic segmentation for Natural Language Processing (NLP). Extraction of word boundary information involves cognitive psychology, computational linguistics, and language education. Methods utilizing word boundary information exist in automatic prosodic and word segmentation, and some have obtained outstanding results. However, the researches of automatic word segmentation in today's NLP are mostly word-oriented, which focus on words instead of word boundaries. As an influential factor in human language recognition, word boundary information is worth more attention and more deeply research.

Keywords

Natural Language Processing; Chinese word segmentation; word boundary information; natural annotation;

1. Introduction

A sentence in any natural language could be considered as a string comprised of words and word boundaries. However, the word boundaries are mostly "invisible" in Chinese language.

[1] This research is supported by National Natural Science Foundation of China, No.60973062, No.61170162.

A few of them are represented as pauses in speeches and punctuations in texts. In raw corpus, word boundary is a kind of objective natural annotation[2].

In different types of corpora, word boundary information exists in different forms and contributes to various applications in many ways. Word boundaries help to form concept boundaries in language acquisition and increase the naturalness of Text to Speech Technology (TTS). Precise recognition of word boundaries benefits disambiguation and improves the accuracy in information extraction. Machine translation could also gain enhancement, since better classification between word boundaries and phrase boundaries helps to narrow the gab between the granularities of source language and target language. All the latter applications in Natural Language Processing (NLP) rely on the utilization of word boundary information in word segmentation.

The recognition of word boundaries in typical forms to segment Chinese strings or utterances, in order to reach proper granularity for more deeply process, is the base of all other techniques in Chinese language processing, which, though has never been completely solved till now.

Benefiting to Chinese word segmentation, word boundary attracts research and interest from cognitive psychology (Amanda Sed and Elizabeth 2006; Katharine Graf Esters et al. 2006; Jusczyk, P. and Aslin, R. 1995; Brent,M. 1999; Li Xing-shan, Liu Ping-ping 2011), language education (Brent, M 1999; Hsu et al. 2000; Xie Hai-yan 2006), and computational linguistics (Jian Zhang et al. 2000; Sun Maosong et al. 2004; Haodi Feng et al. 2003; Hua-Ping Zhang et al. 2010; Zhihui Jin et al. 2006; Tanaka-Ishii 2005; N.Xue 2003; Li Shou-shan, Huang Chu-Ren 2010). How word boundary information helps both infants and adults in language acquisition is discussed in Section 2. Existing methods involving word boundary information in NLP are reviewed in Section 3-5, respectively about prosodic segmentation, word-oriented segmentation, and word-boundary-oriented methods. Section 6 summarizes the mentioned methods and offers our expectations.

2. Word boundary and human language acquisition

In past decades, lingual units of word level acquired more and more important attention in Psycholinguistics (D.W.Carroll 2004). The language-learning children, especially infants, learned word boundary information from edges of utterances and difference of transfer probabilities between various sequences, when they initially start their word segmentation.

[2] Natural annotation is a kind of annotation from users instead of annotators, like user tags, queries, punctuations, which is to some degree equivalent to user generated data. (Sun Maosong 2011)

Experiments demonstrated that, 7.5-month-old infants can recognize sequences "belonging to a single word"(Katharine Graf Estes et al. 2006). Because of pauses, these actions tend to take place near the edges of utterances, especially the end (Amanda Seid and Elizabeth K.Johnson 2006). Stronger ability to segment is found in observation of 17-month-old infants. More boundaries are recognized, such as the two ends of recognized sequences. Sequences with high frequency are more likely to be segmented, which indicates greater sensitivity to different transfer probabilities in utterances. Earlier study (Jusczyk,P. and Aslin, R. 1995) also showed that "words" with higher frequency like names, auxiliaries can help infants cut utterances into smaller units, speeding word segmentation. And the infants in this age are capable of map segmented words to object-labels, which is the initial process of constructing meanings based on statistic segmentation (Katharine Graf Esters et al. 2006).

Clearly, word boundary in forms of pause in speech and transfer probability difference has remarkable functions for infants in segmentation ability, even language acquisition.

Similar conclusions are also found in research on adults' language education, while the word boundaries are in different forms: punctuations, as equivalent counterpart of pauses in speech, become implicit annotations in texts. Study (Ren et al 2010) reported that, the existence of commas reduces fixation time of subjects on key words and sentences, indicating a better understanding of the texts.

In alphabetic writings like English or German, there is another similar natural annotation: the spaces between words. Research in 1960s demonstrated, removing the spaces between words would seriously slow down the reading speed and decrease reading quality (Hochberg et al. 1966).

However, does the word boundary information (like spaces in alphabetic writing) also function to Chinese, Korean or Japanese learners? The answer is positive. In the study of the learners for whom Chinese is learned as second language, when one or half character wide spaces are added between words in texts, increase is found both in reading speed and quality (Brent, M 1999; Hsu et al. 2000), especially in the texts containing high level words and grammars (Hsu et al. 2000).

That is to say, in text reading, word boundaries in various forms, like punctuations or spaces, play a significant role in segmentation and latter semantic process, though some of them are invisible due to historical reasons.

Chinese word segmentation is also a topic in Natural Language Processing. Since it is impossible for lexicon and corpus to cover all lingual phenomena, accumulating of vocabulary, semantic and syntax knowledge are all in serious lack. To some extent, automatic segmentation shares many similar situations with infants' initial segmentation.

Therefore, automatic segmentation, particularly the unsupervised automatic segmentation could also be considered as a simulation of infant segmentation. The relative research will give clues to discover functions of human segmentation, while psychological research can offer new possibility to automatic segmentation technique too (Li Xing-shan, Liu Ping-ping 2011).

Since the word boundaries widely existing in corpora play a so important role in human language acquisition, it is quite natural to fully utilize them in NLP researches.

3. Word boundary and prosodic boundary

In speech, syllables constitute words, and then phrases and sentences. During this process, speaker will insert pauses in different length and pronounce words with various tones and volume. Rhythm of a language comes from this phenomenon, which involves a lot with the naturalness of TTS. In early research, pauses in same length were inserted between each two words, which is against real lingual environment. In order to simulate the real speech, rhythm boundary recognition is in need to predict the length of pauses and other phonetic parameters.

Since phrase is composed linearly by words (Lv Shu-xiang 1979), phrase boundaries are all word boundaries, the former is a subset of the latter. The boundaries of prosodic words are mostly belonging to boundaries of syntactic words, and the boundaries of prosodic phrases are a subset of boundaries of prosodic words. (Zheng Min, Cai Lian-hong 2006) It is obvious that boundary of prosodic phrase is a kind of word boundary. Therefore, the research on prosodic boundary prediction is to some degree incomplete word boundary detection.

A set of features on length of pauses around structural auxiliary words like "的","得","地","所" have been observed by Ying Hong (Ying Hong, Cai Lian-hong 1999), through the analysis of the function word’s characteristics, by studying the phonetic context and position of the structural auxiliary words above in the continuous speech flow and their function in the segmentation of prosodic phrase. Zheng Min (Zheng Min, Cai Lian-hong 2006) builds a two-tier prosodic hierarchy, including prosodic words and prosodic phrases. A statistic model is trained based on the probability frequency of prosodic structure such as part-of-speech, lexical words, length, and position information. An F-value of 88.1% is reached in the test on Tsingua University balanced corpus containing news, novels, essays (56446 characters) and People Daily (130265 characters). The function of single feature part-of-speech is also studied by Niu Zheng-yu (Niu Zheng-yu, Chai Pei-qi 2001).

Part-of-speech information of prosodic boundary context is calculated to train a statistic model to predict the prosodic boundaries in 500 sentences randomly selected from People Daily. Recall rate reaches 60.16% and precision 80.49%.

The later research (Qian Yi-li, Xun En-dong 2006, 2008) takes a simple way by using punctuations in text as natural prosodic annotation corresponding to pauses in speech. In the corpus containing People Daily, Science and Technology Daily, Southern Weekend, Web Pages (825 million characters), all punctuations are replaced with "▲", which is considered as a special word. Trigram statistic model is trained by this corpus after word segmentation. In test, the most likely position of "▲" in a string (sentence at first) would be calculated, and this string is then split at this position, recursively forming a prosodic binary tree.

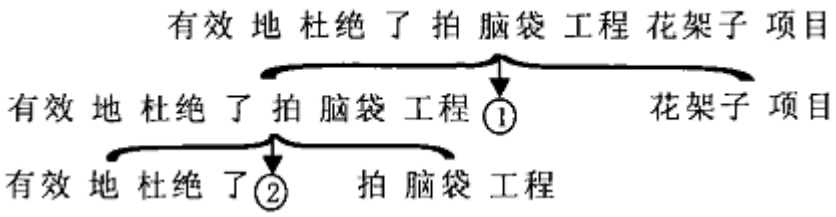

Fig.1 a prosodic binary tree formed by recursively splitting at likely positions of prosodic boundaries

This algorithm is utilized to predict prosodic boundary and obtain prosodic segmented corpus for other research. Though merely one feature is used, Recall rate 83.27% and precision 93.25% are reached in 500 randomly selected sentences (1136 prosodic boundaries) from People Daily under mentioned training data.

4. Automatic word boundary extraction in text

Word boundaries can be indicated by two kinds of knowledge: inner-word knowledge and inter-words knowledge between words. The former is mainly to gain the context information of two edges of a word in order to segment a word from a character flow; the latter describes cohesion in a word, detecting its lower point to find word boundary, though finding word boundary may not be its original purpose. Algorithms (Sun Maosong et al. 2004) combining both kinds of knowledge also exist.

4.1 Inter-Words knowledge based word boundary extraction

Accessor Variety (AV) and Boundary Entropy (BE, or called Context Entropy in some papers) are two main criteria to distinguish words from a character flow.

4.1.1 Accessor Variety

Haodi Feng (Haodi Feng et al. 2004) firstly utilized this criterion in Chinese word segmentation in 2004, with its formula as follow. Here $L_{av}(s)$ is called the left accessor variety, and is defined as the number of distinct characters (predecessors) that precede string s plus the number of distinct sentences of which s appears at the beginning. Similarly, right accessor variety $R_{av}(s)$ is defined. In nature, this criterion describes stability difference between the characters in a word and out of a word. Obviously, the higher accessor variety a string has, in more surrounding contexts it could appear, and inner cohesion in this string is relative stronger: this string is more likely to be a word. This happens to meet one of the traditional linguistic descriptions of Chinese word "stably used" (State Bureau of Technical Supervision 1993).

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\}$$

High-frequency auxiliary words and preposition words (like "是","在","的") and suffix words(like "们") need different processing. After that, Feng extract strings from 1.7 MB Xinhua News. With different thresholds, at most 97.7% of extracted strings turn out to be words.

AV Threshold	Precision	Extracted Strings
2	64.4%	37093
3	83.8%	14468
4	89.6%	8648
5	94.1%	6147
6	96.8%	4757
7	97.4%	3800
8	97.3%	3162
9	97.7%	2734

Table 1. Amount of extracted strings and "be-word" precision under different thresholds

"Stably used" is quite well presented since the "be-word" precision reaches 97.7% when AV threshold is 9. This work impressively demonstrates the function of accessor variety on Chinese word segmentation. However, the recall rate cannot be very high by merely using this criteria since low-frequency and context-dependent words are mostly lost (from AV threshold 2 to 9, over 90% of extracted strings are lost.).

Hua-Ping Zhang (Hua-Ping Zhang et al. 2010) uses N-Shortest Path Methods (Hua-Ping Zhang and Qun Liu 2002) to offer rough segmentation of a sentence. Accessor variety is utilized in analysis of all strings over 2 times, in order to process word out of vocabulary (OOV). Its F-score on Bakeoff 2010 is 95%.

As a recently utilized criterion in Chinese word segmentation, accessor variety has often been used with other features and criteria in statistic models (Hai Zhao and Chunyu Kit 2008; Weiwei Sun et al. 2011).

4.1.2 Boundary Entropy

Similar with accessor variety, boundary entropy is another criterion to describe word boundary (Zhihui Jin et al. 2006; Tanaka-Ishii 2005), which is based on a fundamental linguistic assumption of Harris (S.Z.Harris 1995): when the number of different tokens coming after every prefix of a word marks the maximum value, then the location corresponds to the morpheme boundary. Since information entropy is the measurement of uncertainty, boundary entropy is defined as follow:

$$H(X \mid X_n) = -\sum_{x_n \in \chi_n} P(x_n) \sum_{x \in \chi} P(x \mid x_n) \log P(x \mid x_n)$$

Given a set of elements χ and a set of n-gram sequences χ_n formed of χ, the conditional entropy of an element occurring after an n-gram sequence X_n is defined where P(X=x)indicates the probability of occurrence of x. Some experts found, boundary entropy decreases with pre-words increasing (T.C.Bell et al. 1990). Harries (S.Z.Harris 1995) gave a similar observation on Japanese and Chinese as Fig. 2 shows.

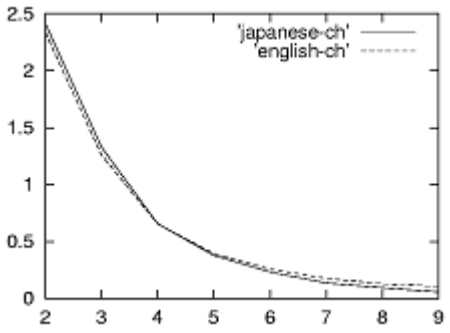

Fig.2.Decrease of character H (X|Xn) when n is increasing

More observation shows that, a local peak indicates a word boundary, as Fig. 3 A. For example, it is easier to guess what comes after "natura" than what comes after "natural" and "nat". Less choices means lower uncertainty, represented as lower entropy. Therefore, surrounding context of a word boundary has lower entropy.

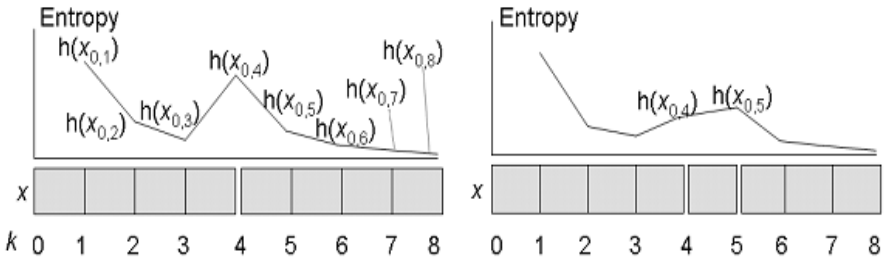

Fig3. Boundary entropy on two words (A) and three words (a single character word in middle, B)

As Fig 3. B, the single character word between two multi-character words also has similar trend on entropy changes.

Ishii (Tanaka-Ishii 2005) trained a statistic model by boundary entropy on 200 MB People Daily, classic literature, and popular magazines. 1MB People Daily is used to test, and precision 90%, recall rate 80% on boundary detection is reached. In Fig. 4, different from the almost linear increase of recall rate, precision stays still on 90% with training corpus enlarged. If more deeply linguistic reasons exist, it will require further research.

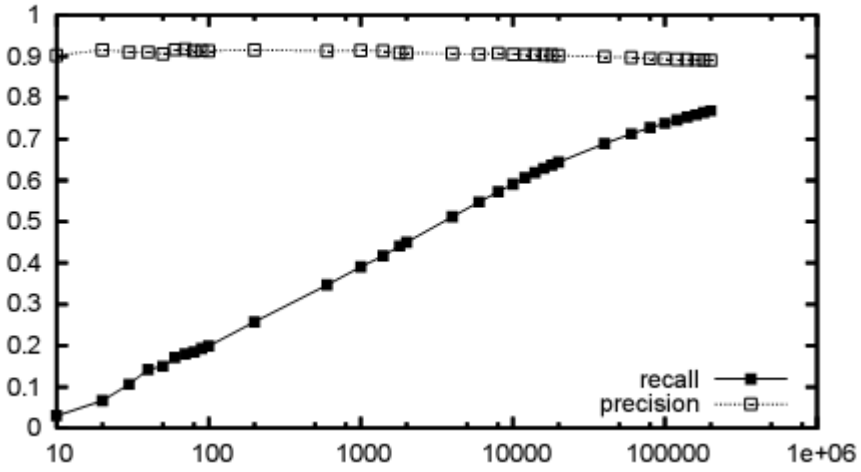

Fig. 4 Precision, recall-rate change with training corpora enlarged

Boundary entropy and accessor variety both describe the uncertainty out of a word. The former is continuous, while the latter is discrete. Chunyu Kit and Zhao Hai reported a better performance from accessor variety than boundary entropy (Hai Zhao and Chunyu Kit 2008, 2011). We believe that, accessor variety is more simple and direct on describing this uncertainty on word boundary, which makes it perform better.

4. 2 Inner-Word knowledge based word boundary extraction

These methods focus on describing the inner cohesion in a string. Mutual Information (MI) is the most often used criteria.

Algorithm based on measuring MI between characters in corpus has been existed since early 1990s (Richard Sproat and Chilin Shih 1990), reaching precision 94% and recall rate 94% on bi-character words. But merely MI-values between each two characters are recorded in calculation, which causes bad performance in multi-character words containing over 3 characters. Its precision decreases to 90% in total.

Relative research (Yang Wen-feng, Li Xing 2001) on 540 MB corpus found, even in bi-character words, 13% of them have MI-value lower than 1.0. Therefore, a great amount of valuable information is lost, when merely MI is used in segmentation.

Zhang Jian integrated Chinese dictionary into the algorithm based on MI, and recorded MI between more characters, in order to solve the problem of long words. Context frequency and word length are all weighted when processing compound words. Over 96%

in precision is reached when extracting compound words from Xinhua News and People Daily in TREC5, TREC6 (7.5 million characters). However, compound word has no clear definition in corpus, and many parameters and thresholds need to be set in length and context weighting. These all limit the effect of algorithm and block the transferring to other type of corpus.

4.3 Word boundary extraction based on "inter & inner" words knowledge

MI is used when deciding if each two characters bound or separated as Fig.5 in Sun Maosong's study (Sun Maosong et al. 2004; M.Sun et al. 1998). If MI-value is over the threshold, two characters are seen as cohesive and bound, then separated.

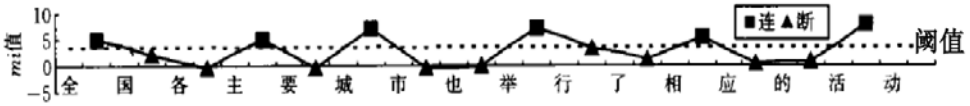

Fig.5 MI-value and character bound decision: squares indicate bound, while deltas mean separated

Based on this frame, Sun developed t-test to different t-test. The former is firstly mentioned by Church in (Church K. W. et al. 1991), measuring to which one a word is more cohesive, when it lies between two other words. Different t-test is defined on string vxyw as follow: (formula 2 is t-test, $\partial^2(p(y|x))$ is variance of $p(y|x)$):

$$dts(x,y) = t_{v,y}(x) - t_{x,w}(y) \qquad (1)$$

$$t_{x,z}(y) = \frac{p(z|y) - p(y|x)}{\sqrt{\sigma^2(p(z|y)) + \sigma^2(p(y|x))}} \qquad (2)$$

MI and different t-test are both weighted in final calculation in decision of "bound-separated". Close test on People Daily of January 1998 reached 85.77% in precision, while 85.88% in open test on People Daily of February 1998.

5. Word boundary oriented word segmentation

No matter inner-word knowledge or inter-words knowledge based methods, word boundaries are not directly described or processed. However all the supervised automatic segmentation methods could be considered as directly describing word boundaries, because complete segmentation information exists already in training corpus. Segmentation based on character tagging is one of the most recent supervised methods and approached outstanding result in evaluations. 5.1 will give an introduction to it. Huang Chu-Ren considers characters as context of word boundaries, focusing on the boundary information to realize word segmentation by classifying different boundaries. This method is discussed in 5.2.

5.1 Automatic Segmentation based on character tagging

Word segmentation can be restated as a tagging task of characters. This idea was firstly introduced by Nianwen Xue (N. Xue 2003) in the beginning of new century. Under the support of Max Entropy Model, traditional 3 tagging method reached over 94% in F-value. Recent utilization of Conditional Random Field (CRF) requires more detailed tagging and description on training corpus: more features from context are integrated. More detailed tagging includes developing 3 tagging into 6tagging. More features like accessor variety are integrated. Zhao Hai approached 95% in F-value on Bakeoff-4 (Zhao Hai and Chunyu Kit 2008) just by integrating AV into training process. Punctuations, AV and MI are all combined in training by Weiwei Sun's recent work (Weiwei Sun and Jia Xu 2011), approaching F-value 96.22% on CTB 6.0. Zhongguo Li reported his study on CRF model with punctuation information, reaching F-value 97.3% in PkU corpus from Bakeoff-2, excluding compound words.

Since segmentation is considered as character tagging, word edges are tagged. Therefore, word boundary information has been acquired in training, and these methods are processing word boundary directly.

5.2 Segmentation based on character boundary classification

Besides punctuations, another natural annotation exits in Chinese texts: character boundaries. Different from Arabic or Persia writings, character boundaries are clear and without ambiguity, due to characteristic of Chinese character. Obviously, word boundaries are a subset of character boundaries. Consequently, research on character boundary will benefit the detection of word boundaries a lot.

Huang Chu-Ren presented the idea in 2007 that, character boundaries could be

considered as targets to process (Huang Chu-Ren et al. 2007). Since word boundaries belong to character boundaries, word boundary detection can be converted into a standard classification task on character boundaries: if character boundary I_i a word boundary B?

$$c_1 I_1, c_2 I_2, ...c_i\ I_i, ...c_{n-1} I_{n-1} c_n$$

A statistic model is trained by recording and calculating all the contexts surrounding "word-word" boundaries. I_i=1 means this character boundary is also a word boundary: otherwise I_i=0. P_{CB}、P_{BC}、P_{CCB}、P_{CBC}、P_{BCC} are calculated as follow.

$$P_{CB}(I_i = 1 | c_i) = \frac{C(c_{i,} I_i = 1)}{C(c_i)}$$

P_{CB} indicates a word boundary that appears after character C_i, and this character boundary is naturally a word boundary. $C(C_i,I_i=1)$ is the count of C_i with a word boundary following; $C(C_i)$ is the count of C_i in corpus. Similarly, P_{CCB} is calculated as follow:

$$P_{CCB}(I_i = 1 | c_{i-1}, c_i) = \frac{C(c_{i-1}, c_i, I_i = 1)}{C(c_{i-1}, c_i)}$$

$C(C_i\text{-}1,C_i,I_i=1)$ indicates the count of a word boundary following $C_{i-1}C_i$. $C(C_{i-1},C_i)$ is the co-occurrences of C_{i-1},C_i in corpus. For the combinations are never been observed in corpus, P value is 0.5, meaning that it is totally uncertain if the following boundary is a word boundary or not. As a result, a 5-dimensional vector is formed to describe the context of each word boundary. Table 2. Shows the vectors of boundaries in string "时间：三月十日".

P_{CCB}	P_{CB}	P_{BC}	P_{CBC}	P_{BCC}	Type	String
0.500	0.595	0.003	0.173	0.021	0	时间
0.983	0.958	1.000	0.998	1.000	1	间：
1.000	0.998	1.000	0.713	0.994	1	：三
0.301	0.539	0.010	0.318	0.054	0	三月
0.964	0.852	1.000	0.426	0.468	1	月十
0.002	0.245	0.065	0.490	0.010	0	十日

Table 2. Feature vectors of boundaries in string"时间：三月十日"

As merely 5 dimensions are used, classifier can be trained by a small corpus. In the later research (Huang Chu-Ren et al. 2008), support vector machine (SVM) turned out to be best in test.

Vectors&Classifi-ers/Corpus	LogReg		LDA		NNet		SVM	
	ASBC	CityU	ASBC	CityU	ASBC	CityU	ASBC	CityU
1,000,000	0.9386	0.9424	0.9325	0.9390	0.9360	0.9423	0.9359	**0.9443**
100,000	0.9389	0.9425	0.9326	0.9387	0.9331	0.9417	0.9369	0.9441
10,000	0.9393	0.9421	0.9326	0.9410	0.9338	0.9409	0.9364	0.9430
1,000	0.9373	0.9419	0.9330	0.9418	0.9334	0.9332	0.9366	0.9400
100	0.9106	0.8857	0.9355	0.9350	0.9198	0.8812	**0.9386**	0.9299

Table 3. Different classifiers and vector amount influent F-value on ASBC and CityU corpus

All the classifiers are trained by Academia Sinaca Balance Corpus (ASBC) in Bakeoff-2. But according to Table. 3, the performance on CityU corpus turns out to be better, demonstrating its good adaptability on different corpora.

Li Shou-shan (Li Shou-shan, Huang Chu-Ren 2010) improved the previous work, in order to conquer the data sparseness in training corpus, by smoothing the calculation of P_{CB} as follow:

$$P_{CB}(I_i = 1 \mid c_i) = \frac{C(c_i, I_i = 1) + 1}{C(c_i) + 2}$$

P_{BC}、P_{CCB}、P_{BCC}、P_{CBC} are modified similarly. Li also incorporates an on-line learning module to acquire new corpus revising recorded data, which is quite meaningful in web era.

6. Summary and future work

As mentioned in Section 1, a sentence could be considered as a string comprising words and word boundaries. So word segmentation is a process to "enlighten" the invisible word boundaries for human readers, for computer, it is carrying out a natural annotation mining task to detect word boundaries.

How could a human reader form a rough segmentation under support of his/her mind

dictionary? Furthermore, how could he/she utilize the boundaries from known words in rough segmentation to build refined segmentation covering all other words? More profound psychological and physical research is required to answer all these questions.

On the other hand, methods in automatic segmentation, both inner-word knowledge and inter-words knowledge based, are all trying to measure a criterion throughout a sentence, detecting its changes and trends to predict a boundary. These methods view word boundary as a same existence with word, ignoring their differences. It may not be a proper solution.

By contrast, Qian Yili and Xun Endong's work directly tags boundaries (Qian Yi-li, Xun En-dong 2006,2008）in Section 3. Huang Chu-Ren et al view word boundary as target, surrounding characters as context, in Section 5 (Huang Chu-Ren et al. 2007, 2008; Li Shou-shan, Huang Chu-Ren 2010). These methods focus on boundary information, facing its difference to word. Though in F-value they cannot beat some integrated methods in Section 3, over 90% in F-value is reached within simple calculation and single feature. Consequently, we believe the methods directly tagging and processing word boundaries have a promising future, and will offer new chances to improve performance of automatic segmentation.

However, two main questions are remained to answer, according to recent practice.

First, do other natural annotations exist besides punctuations and character boundaries?

In today's algorithms and evaluations, numeric strings, quantifier words are often separated to process. Arabic numbers, Latin letters are sometimes not counted in precision and recall rate. Their different characteristics are just shown by this. They and their boundary should be useful to extract and utilize.

It is more valuable, not only to gain more boundary information, but also to narrow the process window for segmentation. According to our statistics in People Daily of January 1998, a pure character string without any punctuation, numeric symbol or Latin letter contains no more than 5 characters in average, which is close to human reading window. Therefore, the word boundary detection will help to reduce segmentation work, better simulating human cognition.

Some characters like "的","着","过", which almost cannot form a multi-character word with other characters, are likely to be indicators splitting a sentence. Since these characters often have some syntactic function, this split may benefit other grammar analysis somehow.

Second, disambiguation or detection on OOV, which problem can word boundary information solve?

The evaluation in terms of Bakeoff data shows that the accuracy drops caused by OOV is 5 times greater than that of segmentation ambiguities (Huang Chang-ning, Zhao Hai 2007). Recent research presented that, word boundary information in form of punctuations may have potential for solving some of the problem, demonstrated by F-value 92.1% in person names and place names, based on a character-tagging model integrating punctuation information.

Person names and place names are indeed a small part of OOV. Which kind of OOV? The new words in true meaning like "朋克" (punk) or compound words like "社会计算" (social computing)? Since no linguistic definition of word is given, the dividing line between compound words and phrase is vague. Would the "word" extracted by boundary information meet our linguistic sense? If word boundary information is really suitable to solve this problem, it would be a very exciting discovery for segmentation research.

More aspects and chances are undoubtedly brought into psychology, language education and NLP by the application of boundary information. With the remarkable rise in the scale of corpus nowadays, implicit annotation (or called natural annotation) mining in raw corpus is becoming increasingly important under the cost pressure of manual annotations. The problems that word boundary information detection is facing require more research in NLP, computational linguistic and other relative subjects to solve.

7. **References**

Amanda Seid and Elizabeth K. Johnson. Infant word segmentation revisited: edge alignment facilitates target extraction.

Brent, M.(1999) . Speech segmentation and word discovery: a computational perspective. Trends in Cognitive Science, 16 (4),298-304.

Church K.W., Gale W., Hanks P., Hindle D ..Using statistics in lexical analysis . In: Zernik U. ed.. Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Hillsdale NJ: Lawrence Erlbaum Associates, 1991, 115~164

D.W. Carroll : Psychology of Language (4th Edition). Shanghai: East China Normal University Press. 2004. P15. (Chinese translated edition)

Fisher, D.f. (1975). Reading and visual search .Memory and Cognition,3,188-196. (Everson's doctoral dissertation)

Hai Zhao, Chunyu Kit: An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJNLP'08) Vol. I, p

9-16, Hyderabad.

Hai Zhao, Chunyu Kit: Integrating unsupervised and supervised word segmentation: The roe of goodness measures. Information Sciences 181 (2011) 163-183.

Hai Zhao, Chunyu Kit: Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition [C]// 6th SIGHAN Workshop on Chinese Language Processing ,Hyderabad, India, 2008: 106-111.

Hanshi Wang, Jiang Zhu, Shiping Tang, Xiaozhong Fan : A New Unsupervised Approach to Word Segmentation. Computational Linguist 2010 Vol.37 Nr.3

Haodi Feng, Kang Chen, Xiaotie Deng, Weimin Zheng : Accessor Variety Criteria for Chinese Word Extraction, Computational Linguistics, 2004,Vol.30 No.1

Hochberg, J.,Levin, H.& Frail,C.(1966). Studies of oral reading: VII: How interword spaces affect reading. Mimeographed, Cornell University. (Everson's doctoral dissertation)

Hsu，S.H.，Huang，K.(2000).Effects of word spacing on reading Chinese text from a video display terminal .Percetual and Motor Skills,90,81 — 92.

Hua-Ping ZHANG,Jian GAO,Qian MO,He-Yan HUANG: Incorporating New Words Detection with Chinese Word Segmentation. In Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010).Beijing, China.2010.8 .p249-251.

Hua-Ping Zhang, Qun Liu: Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method. Journal of Chinese information processing. 2002, 16(5): 1-7

Huang Chang-ning, Zhao Hai: Chinese Word Segmentation: A Decade Review. Journal of Chinese Information Processing.2007 Vol.21, No.3, P8-19.

Huang Chu-Ren, Yo Ting-Shuo, Petr Simon, Hsieh Shu-Kai: A Realistic and Robust Model for Chinese Word Segmentation. Proceedings of ROCLING. 2008

Huang Chu-Ren, Yo Ting-Shuo, Petr Simon, Hsieh Shu-Kai: Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. Proceedings of the Association of Computational Linguistics Annual Meeting (ACL). 2007.

Jian Zhang, Jianfeng Gao, Ming Zhou (2000): Extraction of Chinese Compound Words—An Experimental Study on a Very Large Corpus. Proceedings of the second workshop, P132-139, HongKong, China.

Jusczyk, P.,& Aslin, R. (1995). Infants' detection of sound patterns of words in fluent speech. Cognitive Psychology,29, 1-23

Katharine Graf Estes, Julia L. Evans, Martha W. Alibali, and Jenny R. Saffran . Can Infants Map Meaning to Newly Segmented Words? Statistical Segmentation and Word

Learning.

Li Shou-shan, Huang Chu-Ren: Chinese Word Segmentation Based on Word Boundary Decision. Journal of Chinese Information Processing. 2010, Vol. 24, No. 1, P3-7.

Li Xing-shan, Liu Ping-ping, Ma Guo-jie: Advances in Cognitive Mechanisms of Word Segmentation During Chinese Reading. 2011,Vol.19, No.4, 459-470.

Lv Shu-xiang: Chinese Grammar Analysis, Beijing: Commercial Press, 1979, P39,P30.

M. Sun, D. Shen, and B. K. Tsou. 1998 . Chinese word segmentation without suing lexicon and hand-crafted training data. In COLING-ACL.

N.Xue: Chinese Word Segmentation as Character Tagging [J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1), 29-48

Niu Zheng-yu, Chai Pei-qi: A Statistical Approach Based on Boundary POS Feature to Prosodic Phrasing. Journal of Chinese Information Processing. 2001, Vol. 15, No. 5, P19-25.

Qian Yi-li, Xun En-dong: Prediction of Speech Pauses Based on Punctuation Information and Statistical Language Model. PR&AI. 2008, Vol. 21, No. 4, P541-545.

Ren ,G. ,&Yang ,Y.(2010).Syntactic boundaries and comma placement during silent reading of Chinese text: evidence from eye movements .Journal of Research in Reading,33,168–177.

Richard Sproat, Chilin Shih (1990): A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages Vol.4 No.4

S.Z. Harris. From phoneme to morpheme .Language , 1995, P 190-222.

State Bureau of Technical Supervision. The People's Republic national standards GB/T13715-92 Contemporary Chinese Language Word Segmentation Specification for Information Processing. Beijing: Standards Press of China, 1993.

Sun Maosong, Xiao Ming, Tsou B K: Chinese Word Segmentation without Using Dictionary Based on Unsupervised Learning Strategy. Chinese Journal of Computers. 2004, Vol. 27, No. 6, P736-742 .

Sun Maosong: Natural Language Processing Based on Naturally Annotated Web Resources. Journal of Chinese Information Processing. 2011, Vol. 25, No. 6, P26-32.

T.C. Bell, J.G. Cleary, and Witten. I.H .1990 .Text Compression .Prentice Hall.

Tanaka-Ishii .2005 .Entropy as an indicator of context boundaries ——an experiment using a web search engine. In IJCNLP, P 93-105.

Weiwei Sun, Jia Xu: Enhancing Chinese Word Segmentation Using Unlabeled Data. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, P 970-979.

Xie Haiyan: Research on the Chinese Word Boundary Parsing Ability of Intermediate and

Advanced Foreign Students. Jinan University, Masters Thesis 2006.6.

Xun En-dong, Qian Yi-li, Guo Qing, Song Rou: Using Binary Tree as Pruning Strategy to Identify Prosodic Phrase Breaks. Journal of Chinese Information Processing. 2006, Vol. 19, No. 3.

Yang Wen-feng, Li Xing: PAT-TREE Based Language Model and Automatic Keyword Extraction. Computer Engineering and Applications. 2001, Vol. 37, No. 1, P16-19.

Ying Hong, Cai Lian-hong: Research on the Segmentation of the Prosodic Phrase Based on Driven by the Structural Auxiliary Word. Journal of Chinese Information Processing. 1999, Vol. 13, No. 6. P41-46.

Zhihui Jin, Kumiko Tanaka-Ishii (2006): Unsupervised Segmentation of Chinese Text by Use of Branching Entropy. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, P428-435

Zheng Min, Cai Lian-hong: Statistical Model Based on Probability Frequency for Mandarin Prosodic Structure Prediction. J Tsinghua Univ (Sci & Tech), 2006,Vol. 46, No. 1, P78-81.

Zhongguo Li, Maosong Sun : Punctuation as Implicit Annotations for Chinese Word Segmentation. Computational Linguist. 2009 Vol.35 Nr.4.