Chinese Natural Chunk Research Based on Natural Annotations in Massive Scale Corpora[*]

Exploring Work on Natural Chunk Recognition Using Explicit Boundary Indicators

Zhi-e Huang, En-dong Xun, Gao-qi Rao, and Dong Yu

BLCU, International R&D Center for Chinese Education
{hze_blcu,raogaoqi_fj}@163.com, {edxun,yudong_blcu}@126.com

Abstract. Great changes in Natural Language Processing (NLP) research appear with the rapid inflation of corpora scale. NLP based on massive scale natural annotations has become a new research hotspot. We summarized the state of art in NLP based on massive scale natural annotated resource, and proposed a new concept of "Natural Chunk". In the paper, we analyzed its concept and properties, and conducted experiments on natural chunk recognition, which exhibit the feasibility of natural chunk recognition based on natural annotations. Chinese natural chunk research, as a new research direction in language boundary recognition, has positive influences in Chinese computing and promising future.

Keywords: Massive scale corpora, Natural Annotation, Natural Chunk, Natural Chunk recognition.

1 Introduction

Language boundary plays a significant role in human language acquisition. Language boundary is a basic explicit feature of language unit, on which research is known as a basic research in linguistic [1]. In Chinese, different terms are used in description of the concept of "word", like "syllable word", "separable word", "short phrase", "rhythm word", "rhythm phrase" etc [2]. Different inner sentence units are defined for various applications in Natural Language Processing (NLP), like Chinese word segmentation, chunking, rhythm phrase recognition etc, which also require different computing methods.

With the development of Internet, as to corpora construction, data scale is no longer the main difficulty. With the improvement of hardware performance and distributed computing technology, algorithm barriers decrease a lot. All these influence the NLP research and attract considerable attentions to exploring methods of automatically acquiring linguistic knowledge from raw corpus. [Sun, 2011] proposed the concept of "NLP based on massive scale natural annotated corpora" and stated that Natural

[*] Supported by NFSC(61170162), State Language Commission (YB125-42), National Science-technology Support Plan Projects (2012BAH16F00) and the Fundamental Research Funds for the Central Universities(13YCX192).

M. Sun et al. (Eds.): CCL and NLP-NABD 2013, LNAI 8202, pp. 13–24, 2013.

Language Processing Based on Huge-scale Naturally Annotated Corpora should be the future direction of NLP research [3].

A natural sentence is made up of words, among which exist cohesion and transition relations. In massive scale of corpora, these relations could be relieved by natural annotations. Theory of Prefabricated Chunk, rooted in cognitive linguistic, states the existence of lots of directly used poly-word strings with solidification and prefabrication. Combining the theory of Prefabricated Chunk with Natural Annotation oriented NLP research, we put forward a new language unit, define the solidified strings that frequent and stably appear in various contexts as 'Natural Chunks'.

Natural chunk is a kind of language unit observed from pragmatic level. Natural chunk recognition aims to meet various application requirements on language boundaries within a united framework. Boundaries' strength should be maintained, so as to support the further recognition of other language units' boundaries among a sentence. The task of natural chunk recognition is in fact the word boundary prediction based natural annotation in massive scale of corpora. This task contains 3 parts, namely natural annotation mining, natural chunk boundary modeling, and evaluation. By parameters modification, different chunking for one sentence could be approached. Existed relative work shows that chunking based on natural annotation is feasible, and has positive influences in Chinese computing.

2 Language Computing Based on Natural Annotated Boundary Knowledge

Natural annotation resource consists of various resource data generated by users in all means, like web pages, forums, twitters, Wikipedia, etc. In the view of natural language processing (NLP), these data can be seen as partially tagged. And the natural annotations are quite worthy to be utilized.

On its application in NLP, first, massive scale is the necessary condition to make Natural Annotation computing into play. Second, shallow processing is the basic tone of calculation on massive scale corpora. At last, professional linguist knowledge should not be refused in any case in analysis and processing [3]. Relative researches in Chinese word segmentation, rhythm phrase recognition, information extraction, social computing, sentiment analysis, text classification and information retrieval have came to appear in recent years.

Natural annotations are rich of lexical information. Experimental and research results indicate that they can benefit language boundary recognition. Language boundary recognition researches took advantages of natural annotations are mainly about Chinese word segmentation and rhythm phrase recognition. [Rao etc,.2013] explored the lexicon knowledge containing by punctuations, Latin letters and Arabic number in small data set [4]. [Zhongguo Li etc,. 2009] used punctuation as segmentation features offering positive cases, increasing word segmentation performance of CRF model, especially for OOV (words Out Of Vocabulary) in BakeOff test [5]. [Yuhang Yang etc,. 2008] research about Chinese term extraction based on delimiter [6]. [Xing Li etc,. 2006] introduced punctuations as a important feature into hierarchical Chinese chunking [7].

[Thomas etc,. 2005] increased precision on Chinese-English sentence pair alignment by adding punctuation as feature [8]. [Qian Yili etc,. 2008, Xun Endong etc,. 2005] proposed method of building rhyme structure binary tree, based on the feature of punctuation location. High precision was reached with low training cost [9,10]. [Valentin etc., 2010] increased unsupervised chunking 5 percent with features of HTML tags (anchor, bold, italic and underline). Punctuations were also used in chunking [11]. [Valentin etc,. 2011; Weiwei Sun etc., 2011] merged punctuations into other traditional statistic features, resulting to better performance in OOV [12,13].

In brief, Natural annotations are unconsciously annotated by users, which avoid the problem of tagging cost. With massive scale corpora, both explicit and implicit annotation mining will partially conquer the difficulty in quality assurance and quantity limit, benefitting various tasks in NLP.

3 Chinese “Natural Chunk”

Theory of Prefabricated Chunk (PC), rooted in cognitive linguistic, states the existence of lots of directly used poly-word strings with solidification and prefabrication. Prefabrication indicates that syntax generation and inner syntax analysis are not in need in the usage of PC each time. A natural sentence is made up of words, among which exist hierarchy and cohesion relations. With a massive scale of corpora, these properties could be relieved by the natural annotations which richly contain boundary information, appear in a specific form of language unit. This language unit is defined as “Natural Chunk”.

Definition. The language units that continuous stable and frequent appear with distinctive boundary features in massive scale corpus are ‘Natural Chunks’.

In Chinese, natural annotations like punctuations, Arabic numbers, Latin letters match the definition of natural chunk, and they are special ones.

Properties of “Natural Chunks”

(a) Integrity (inner cohesion). A natural chunk is a continuous string. Conventionalized usages are often seen as natural chunks like “与此同时(at the same time)”, “也就是说(that is to say)”

(b) Stablilty. In massive scale corpora, frequently used in various contexts.

(c) Boundary Features. Natural chunks is strings with distinct boundaries. The boundary information offered by natural annotations in massive scale corpora. Different from other language units which are bounded with syntactical rules.

(d) Application Oriented. The properties of natural chunks depend on the properties of the natural annotations used in the recognition progress. For various applications, the mining and usage of various natural annotations adopting in the process varied, different but rational natural chunk recognition can be obtained. Moreover, their evaluations should also adapt to the specific application.

Properties of “integrity” and “stability” are quite similar to the principles of Chinese word segmentation. If punctuations are used as Natural Annotation in recognition, chunks also often match rhyme structures in one sentence. Natural Chunking has advantages in Chinese language boundary computation, for it is not bound with syntactical rules as the others do, and also due to its unsupervised mining process in massive scale corpora.

4 Natural Chunk Recognition in Massive Scale Corpora

The work in this section is consisted by three parts, namely the mining of natural annotations with distinctive boundary information, the boundary modeling and evaluation. All these work are concentrating on the boundary computing of natural chunks.

4.1 Natural Annotations with Distinctive Boundary Information

We use BIC (Boundary Information Carrier) to signify natural annotations carrying distinctive boundary information expressed. Base on whether a BIC is intuitive and easy to extract, classify BICs into explicit BICs and implicit BICs. A explicit BIC should be intuitive and easy to extract. To be specific, Punctuation, line break, Arabic numbers and Latin letters are explicit BICs, because they are easy to be identified and extracted from Chinese sentences, since they do not belong to Chinese character set and never associate with other Chinese characters as a word. Ex1 present how punctuations and Arabic numbers separate the chunk out of a sentence. “的” (of), “与” (and) and “对垒” (confront) are implicit BICs. By the way, using explicit BICs to acquire implicit BICs, is one of our studies in the future.

Ex.1

(a) …… 。 改革开放以来 ，老百姓开矿治穷……
(b) “ 站住 ！” 值班的战士大吼一声……
(c) ……截止收盘沪指报2293.08点，涨4.55点……
(d) 在市场上站住脚 与 站稳脚 的 思路同样有差异。
(e) 她们将同老对手、实力强劲的[org] 对垒 ……

As subset of natural annotations, BICs contain rich linguistic knowledge, including shallow formats and pragmatic patterns etc. High coverage of various lingua phenomena could be approached. By iterations, more BICs especially implicit ones could be gained. Ex.3 shows how to gain a chunk “篡改财务账目” (tampering with financial accounts) in d) from “改革开放” (reform and opening up) in a) by “特别是 ”(especially) in b) and “通过采取” (by taking) in c). BIC mining, boundary strength computing and iteration strategy is critical in this part.

Ex.2

(a) 罗干说 ， 改革开放以来 ， 中国旅游业……
(b) 。特别是←改革开放以来 ， 人民子弟兵……
(c) 特别是→通过采取 “切割”、“站票”……
(d) 。通过采取→篡改财务账目 的 方式贪污赃款

4.2 Natural Chunk Boundary Modeling

Natural chunks are strings with distinctive boundaries. It is crucial to utilize context information in boundary modeling. On the other hand, a natural chunk is stably and frequently used in massive scale corpora. Its inner cohesion is another influencing factor of boundary modeling. A natural chunk's boundary could be gained at the low point of inner cohesion. In brief, both inner cohesion and autonomy in context are import features in boundary modeling.

Similar with Chinese word segmentation, frequency, mutual information (MI), boundary entropy (BE) and accessor variety (AI) are often used to describe the cohesion and isolation of a "word" [14]. [Hanshi Wang etc, 2011] proposed algorithm of "Evaluation-Segmentation-Adaption" merging boundary entropy, frequency and length, and obtained a segmentation after a convergence iteration process [15].

Natural chunks are flexible in granularity. However within a specific application, a sentence has only one specific and rational natural chunk recognition result, matching the needs of various applications by parameter adaption. Boundary strength is necessary information in natural chunk recognition. Any strings in a sentence, its isolation to be a natural chunk could also be described by boundary strength in boundary modeling. We will investigate the cohesion and isolation of natural chunk and its boundary modeling combining both.

4.3 Evaluation of Natural Chunk Recognition

Natural chunk recognition, like Chinese word segmentation as well as rhythm recognition, is a kind of research on language boundary identification. But it should be pointed out that natural chunk recognition aims to gain a united framework for various kinds of language boundary recognition. Within the framework, by modification of parameters and decoding strategies, different needs of applications in language boundaries should be matched. And further evaluations should also fit the application needs.

To make it simple, Chinese lexicon is in natural small natural chunks solidified in pragmatics, while rhythm phrase can be viewed as natural chunks in coarse granularity. From this point, fine natural chunking of small granularity for segmentation while coarse granularity for phrase recognition.

If Chinese word segmentation is viewed as application context, parameter tuning can be realized comparing with tagged corpus, and made its chunking result close to the granularity of word. Its test result would also be evaluated with the standard of word segmentation. Similarly is the application context of rhyme phrase.

5 Experiments

5.1 Corpus and Dataset

In this paper, the corpus in use is a massive scale balanced monolingual corpus, consisted of Beijing Language and Cultural University International R&D Centre for Chinese Education. The corpus contains News (People Daily), Literature, Weibo (Chinese Tweeter, 3 months data) and Blog (3 months data), of total size 102.8 GB (ASCII). Size of each part is shown in Table.1.

Table 1. Data Distribution of each Part of Corpus

Type	News	Literature	Weibo	Blog	Total
Scale (B, ASCII)	6.26G	8.29G	35.0G	53.3G	102.8G

5.2 Lexicon Knowledge in BICs

Punctuations are special natural chunks. In Chinese, punctuation is a close set and easy to recognize, carrying distinctive boundary information. All punctuations in the previously mentioned massive scale corpus are replaced by "∆", and used as a segmentations' symbol and segmented the corpus.

The Modern Chinese Dictionary (5th version), 'MCD5' for short, contains 62777 different Chinese word types. Substrings of 1-3 characters and 4-8 characters are extracted from the segmented corpus mentioned above. Among them, over 96% words of MCD5 have been covered. In fact, [Rao Gaoqi etc, 2013] reported in the small data set of 350MB (ASCII), 87.84% words can be covered. Fig.1 shows the increasing of word coverage with pruning frequency[4]. By then, we can come to an conclusion that explicit Natural annotations like punctuation in big data set richly contain language boundary knowledge.

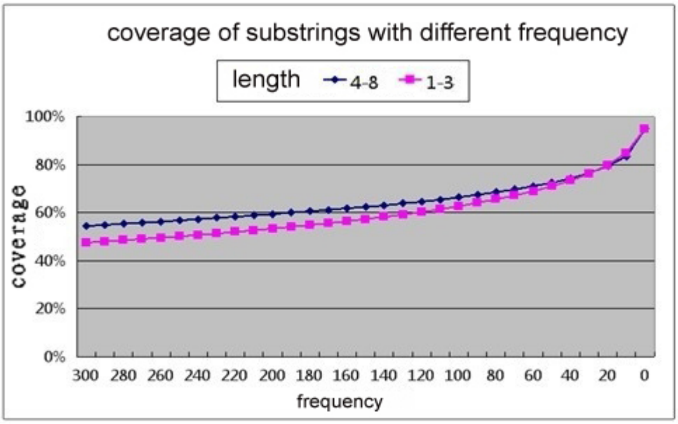

Fig. 1. Words Coverage in Modern Chinese Dictionary (5th version)

5.3 Word Segmentation Experiments with explicit BICs

A string $C_1C_2\cdots C_{i-1}C_iC_{i+1}\cdots C_{n-1}C_n$ can be equivalent to an interval form as $C_1I_1C_2I_2\cdots C_{i-1}\ Ii\text{-}_1C_iI_iC_{i+1}I_{i+1}\cdots C_{n-1}I_{n-1}C_nI_n$. In interval form, I_i=1, when it is a word boundary (*B* for short) or I_i=0. Straightforwardly p(B_i)=p(I_i=1) [16,17].

Assumption 1. Boundary strength (statistically indicates the probability for a character boundary being a word boundary) is only relative to the adjacent context. Therefore in massive corpus, the boundary information carried in BICs could highly cover various language boundaries.

"Δ" stands for punctuation. We present p(B_i) in tri-gram form. And we have:

$$\begin{aligned} p(B_i) &= p(C_{i-1}C_iI_i = 1C_{i+1}C_{i+2}) \\ &= p(C_{i-1}C_i, I_i = 1) \times p(C_i, I_i = 1, C_{i+1}) \times p(I_i = 1, C_{i+1}C_{i+2}) \\ &\approx p(C_{i-1}C_i\ \Delta) \times p((C_i\ \Delta\ C_{i+1}) \times p(\Delta\ C_{i+1}C_{i+2}) \end{aligned} \quad (1)$$

Assumption 2. Boundary strength is positive correlated to the isolation of its context, while negative correlated to the cohesion of its context.

According to the assumption that BICs richly contains boundary knowledge, considering positive correlation exists between boundary strength and its co-occurrence frequency with the punctuations, while negative correlation exists between boundary strength and its co-occurrence frequency of context in a massive scale corpus. Hence having f(B_i) formulated as (4) in a context window of width 4 characters, within which "Δ" stands for any punctuation.

$$f(B_i) = \frac{C(C_{i-1}C_i\ \Delta) \times C(\Delta\ C_{i+1}C_{i+2})}{C(C_{i-1}C_iC_{i+1}C_{i+2})} \quad (2)$$

Test set1 and 2 are extracted from People's Daily (Jan. 1998). 5328 sentences from January (19.1 characters in average length) and 5328 sentences from December (19.3 characters in average length).

Table 2. Test Sets

Test Set	Character Number	Character Type	Word Number	Average word Length
Set1	67896	2392	39337	1.7260086
Set2	68424	2261	39803	1.7190664

Tuning point distinction is chosen as decoding strategy. For each character boundary I_i, comparing its previous boundary I_{i-1} and following boundary I_{i+1}, if p(I_i)⩾p(I_{i-1}) (or, f(I_i)⩾f(I_{i-1})) and p(I_i)⩾p(I_{i+1}) (or, f(I_i)⩾f(I_{i-1})), then *I*i=1 and *I*i is a word boundary *B*i; or *I*i=0 and *I*i is not a word boundary.

Precision (P in short), Recall (R in short)and F-0.5 were used for evaluation. F-0.5 value combines precision and recall, and it emphasizes precision. They are defined as formula (3), (4) and (5). *A* denotes word boundaries by manual annotated, while N_A is set *A*'s count. *B* denotes word boundaries tagged by our algorithm, and N_B is the count of set *B*.

Why not F1? For the natural chunk recognition's result – natural chunks might be somehow rough than the words. As a string "AB" is stably and frequently appear with BICs in a massive scale corpora, according to definition of natural chunk, we might take "AB" as a natural chunk, however in segmentation, it might be take in the form of "A" and "B". Since we are more interested in natural chunk recognition than rarely Chinese word segmentation, we'd prefer to take F-0.5 value than F-1.

$$P = \frac{A \cap B}{N_A} \times 100\%, \tag{3}$$

$$R = \frac{A \cap B}{N_B} \times 100\% , \tag{4}$$

$$F(0.5) = \frac{(1+0.5^2)*P*R}{0.5^2*P+R} \times 100\% \tag{5}$$

BE, AV and MI are often used statistic features. We build a baseline system, based on character entropy. Means of left entropy and right entropy are used as description of boundary strength.

Table 3. Performance of Baseline System

Test Set	Type	Avg Len	P	R	F-1	F-0.5
Set1	Blog	2. 5988	75. 64%	50. 24%	60. 37%	68. 69%
	Literature	2. 6040	67. 26%	44. 58%	53. 62%	61. 05%
	News	2. 5663	76. 80%	51. 65%	61. 76%	69. 98%
	Weibo	2. 5376	64. 91%	44. 15%	52. 55%	59. 33%
Set2	Blog	2. 5995	78. 39%	51. 84%	62. 41%	71. 11%
	Literature	2. 6177	69. 77%	45. 82%	55. 32%	63. 17%
	News	2. 5626	79. 06%	53. 04%	63. 49%	72. 00%
	Weibo	2. 5766	66. 18%	44. 15%	52. 97%	60. 17%

5.4 Results and Analysis

Table.4 presents the result based on assumption 1, and Table.5 shows the result of experiments based on assumption 2. it is easy to observe that just by the boundary information contained in explicit BICs, tri-gram method could beat baseline system (nearly 18 percentages increased in both precision and F-0.5). Even if we alternate the corpus resource to Sina Blog (different type of writing to People's Daily), this advantage changes little. By about 3 percentages in precision and 1 percentage in F-0.5 increases.

Mentioned results suggest the effectiveness of the natural boundary computing in massive scale of corpus. It is also easy to observe that stylistic difference influences little on boundary recognize, that will benefit cross domain research.

Table 4. Results of Tri-gram ans Assumption 1 Model

Test Set	Type	Avg Len	P	R	F-1	F-0.5
Set1	Blog	2. 197566	93. 42%	73. 38%	82. 20%	88. 58%
	News	2. 205848	93. 34%	73. 03%	81. 95%	88. 42%
Set2	Blog	2. 190479	93. 57%	73. 44%	82. 29%	88. 71%
	News	2. 203955	93. 57%	72. 99%	82. 01%	88. 58%

Table 5. Results of Assumption 2 Model

Test Set	Type	Avg Len	P	R	F-1	F-0.5
Set1	Blog	2. 4142	96. 59%	69. 06%	80. 54%	89. 46%
	Literature	2. 4363	95. 40%	67. 58%	79. 12%	88. 14%
	News	2. 4137	96. 57%	69. 05%	80. 53%	89. 44%
	Weibo	2. 4352	95. 98%	68. 03%	79. 62%	88. 69%
	Mixed (all)	2. 4202	96. 51%	68. 83%	80. 35%	89. 32%
Set2	Blog	2. 4051	96. 68%	69. 11%	80. 60%	89. 54%
	Literature	2. 4366	95. 87%	67. 64%	79. 31%	88. 48%
	News	2. 4025	96. 57%	69. 10%	80. 56%	89. 46%
	Weibo	2. 4199	96. 01%	68. 20%	79. 75%	88. 77%
	Mixed (all)	2. 4104	96. 67%	68. 94%	80. 48%	89. 47%

If we valued 1 for the word boundary in golden standards and 0 for none word boundary. Boundary scoring is normalized by the division of the scoring of the whole sentence. Fig.2 presents the boundary scoring of sentence “迈向充满希望的新世纪” (Towards a new century which is full of hope) and its manual segmentation. An optimal segment point could be found, and sentence will be split into two. Recursively processed by the same steps until the segmented strings match the expectation as a “word”. Fig.3 is the binary segment tree formed in recursive process on the sentence. We could find strong isomorphism of our result with word segmentation standard, and in some levels of binary segment tree, chunk boundaries can also match the rhyme structure of this sentence.

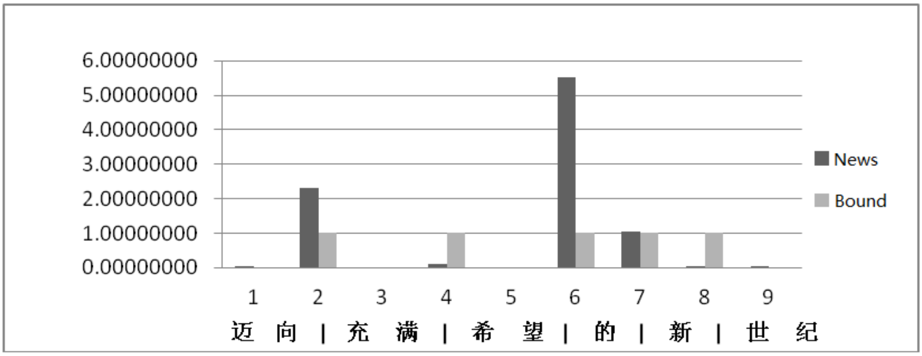

Fig. 2. Boundary Segmentation and Manual Segmentation

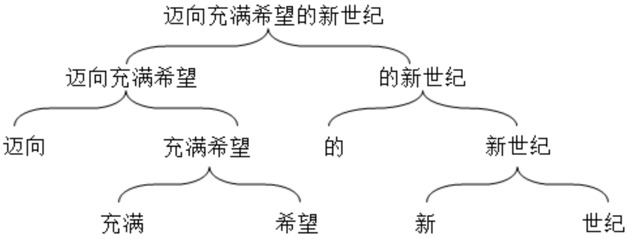

Fig. 3. Binary Segment Tree

Ex.3 indicates the good performance in named entity recognition, especially in long sentences, using punctuations and implicit BICs. "与" (and), "为" (for), "和" (and), "的" (of) etc., are characters (single character word) with strong boundary strength. They are quite worthy to be used in boundary computing. As for future work, it would be natural for researchers to enhance the implicit Natural Annotation utilization as well as modeling both cohesion and autonomy of a string.

Ex.3 Segmentation Examples

(a) 江泽民 李鹏 乔石 朱镕基 李瑞环 刘华清 尉健行 李岚清 与万名 首都 各界 群众 和劳动 模范 代表 一起 辞旧迎新
(b) 本报讯 广东鹤山市 直达 香港 九龙的 豪华客车专线 日前开 通
(c) 恭城瑶族自治县 提供了 成功 的经验
(d) 刚刚在 英国 首都 伦敦 为争取 英国 政府 释放 智利 前总统 皮诺切特 进行了 三天 游说 活动的 智利 外长 因苏尔萨 将于 n日 赶赴 西班牙 首都 马德里

6 Summary and Future Work

A natural sentence is made up of words, among which exist cohesion and transition relations. In massive scale of corpora, these relations could be relieved by utilizing the co-occurrence with natural annotations in massive scale corpora. Theory of Prefabricated Chunk (PC), rooted in cognitive linguistic, states the existence of lots of directly used poly-word strings with solidification and prefabrication. Combining this phenomena and natural annotation oriented NLP research, we define that the "Natural Chunk" is the solidified, frequent string that stably collocate with various contexts.

Natural chunks are language units observed from pragmatic level. Natural chunk recognition aims to meet various application requirements on language boundaries in a united framework. Information of boundary strength should be contained, so that the further description on different boundaries from characters to sentence could be possible. The task of Natural Chunk recognition is in fact the word boundary prediction based natural annotation in massive scale of corpora. This task contains 3 parts, namely natural annotation mining, chunk boundary modeling, and chunking evaluation. By parameters modification, different chunking for one sentence could be approached. Existed work shows that chunking based on natural annotation is quite effective and has promising future.

As for boundary prediction based on massive scale of corpora, it is a new task serving other relative applications. It is worth noting that, current work still ongoing. Boundary modeling combining isolation and inner cohesion, features selection and pruning criteria are all worthy task waiting to research.

References

1. Liu, C.: Structure and Boundary - A Cognitive Study on Linguistic Expressions. Shanghai Foreign Language Education Press (December 2008)
2. Feng, S.: The multidimensional properties of "word" in Chinese. Contemporary Linguistics 3(3), 161–174 (2001)
3. Sun, M.: Natural Language Processing Based on Naturally Annotated Web Resources. Journal of Chinese Information Processing 25(6), 26–32 (2011)
4. Rao, G., Xun, E.: Word Boundary and Chinese Word Segmentaion. Journal of Beijing University (Natural Science Edition) 49(1) (2013)
5. Li, Z., Sun, M.: Punctuation as implicit annotations for Chinese word segmentation. Computational Linguistics 35(4), 505–512 (2009)
6. Yang, Y., Lu, Q., Zhao, T.: Chinese Term Extraction Based on Delimiters. In: Conference: Language Resources and Evaluation – LREC (2008)
7. Li, X., Zong, C.: A Hierarchical Parsing Approach with Punctuation Processing for Long Chinese Sentences. Journal of Chinese Information Processing 20(4), 8–15 (2006)
8. Chuang, T.C., Yeh, K.C.: Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria. Computational Linguistics and Chinese Language Processing 10(1), 95–122 (2005)
9. Qian, Y.-L., Xun, E.-D.: Prediction of Speech Pauses Based on Punctuation Information and Statistical Language Model. PR&AI 21(4), 541–545 (2008)

10. Xun, E.-D., Qian, Y.-L., Guo, Q., Song, R.: Using Binary Tree as Pruning Strategy to identify Rhythm Phrase Breaks. Journal of Chinese Information Processing 20(3), 23–28 (2006)
11. Spitkovsky, V.I., Jurafsky, D.: Profiting from mark-up: Hypertext annotations for guided parsing. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1278–1287 (2010)
12. Spitkovsky, V.I., Alshawi, H., Jurafsky, D.: Punctuation: Making a Point in Unsupervised Dependency Parsing. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 19–28 (2011)
13. Sun, W., Xu, J.: Enhancing Chinese Word Segmentation Using Unlabeled Data. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 970–979 (2011)
14. Zhao, H., Kit, C.: An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. In: International Joint Conference on Natural Language Processing – IJCNLP 2008 (2008)
15. Wang, H., Zhu, J., Tang, S., Fan, X.: A New Unsupervised Approach to Word Segmentation. ACL 37(3), 421–454 (2011)
16. Huan, C.-R., Šimon, P., Hsieh, S.-K., Prévot, L.: Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. In: Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 69–72 (2007)
17. Li, S., Huang, C.-R.: Chinese Word Segmentation Based on Word Boundary Decision. Journal of Chinese Information Processing 24(1), 3–7 (2010)