Word segmentation for computational oracle bone inscriptions

Jing Xiong

School of Computer and Information Engineering,
Anyang Normal University, Anyang 455000, Henan, China

Xiaoqi Niu

School of Mathematics and Statistics, Anyang Normal University,
Anyang 455000, Henan, China

[*]Yihua Lan

School of Computer and Information Technology,
Nanyang Normal University, Nanyang 473061, Henan, China
*Corresponding author: Yihua Lan

Abstract
According to Oracle Bone Inscriptions (OBI)'s own characteristics, a word segmentation method based on the combination of dictionary, word frequency, part of speech, sliding window segmentation algorithm is proposed. Firstly, an OBI dictionary is built to support word segmentation; secondly, a sliding window algorithm is introduced; finally, the word sense disambiguation and context analysis problem related the word segmentation result based on sliding window algorithm and OBI dictionary are proposed. The proposed method can largely remove ambiguities and improve the identification of OBI word segmentation. Experimental results show that the method is according to the grammatical features of OBI, and it can obtain higher accuracy and recall than any other word segmentation method or tool which designed for modern Chinese.
Key words: COMPUTATIONAL ORACLE BONE INSCRIPTIONS, WORD SEGMENTATION, WORD SENSE DISAMBIGUATION, SLIDING WINDOW

Introduction

OBI has more than 3,500 years' history. It record a wide range of social activities of the Shang Dynasty royal divination, including royal historical events, agricultural, astronomical phenomena, sacrifice, conquest, imperative, exchanges and marriage. So it has very rich historical content and important research value. OBI was first discovered in 1899, after more than 115 years of development, the research related OBI has formed an international eminent study with strict rules, abundant research material and

multidisciplinary. OBI is closely related to some other subjects including philology, history, archaeology, ancient history of science, history of literature and anthropology [1].

The traditional OBI research methods are of great difficulty. For an OBI expert need ten or twenty years or even longer time, which seriously hindered the progress of OBI research. Study on the improvement of the traditional way of using information technology, general linguistics, logic, philosophy, computer science, artificial intelligence, mathematics and statistics and other disciplines to study the OBI information processing can solve this problem, which we named Computational Oracle Bone Inscriptions.

Due to the rapid development of computer science and information technology, it provides a very convenient condition to research the ancient Chinese deeply. With the development of ancient Chinese characters study, it is urgent to take advantage of more complete and comprehensive information on the original word OBI. And it is particularly important to use computer to realize the OBI recognition and interpretation automatically. Among them, OBI word segmentation is the first and important step for computer processing.

Currently, there are many researches about Chinese word segmentation and some of them have gained high achievement. They can be divided into three categories: word segmentation method based on dictionary and thesaurus; word segmentation method based on word frequency statistics; word segmentation method based on knowledge understanding. Reference [2] analyzed the Chinese Word Segmentation papers published from 2004 to 2008 by using the principles and methods of bibliometrics. It discussed the current research situation of Chinese word segmentation in China through the author analysis and the distribution of literature on Chinese word segmentation. Reference [3] present a Chinese word segmentation system which was built using a conditional random field sequence model that provides a framework to use a large number of linguistic features such as character identity, morphological and character reduplication features. Reference [4] proposed a method that effectively combines the strength of both segmentation schemes using an efficient dual-decomposition algorithm for joint inference. The method is simple and easy to implement. Reference [5] present a novel lattice-based framework in which a Chinese sentence is first segmented into a word lattice, and then a lattice-based POS tagger and a lattice-based parser are used to process the lattice from two different viewpoints: sequential POS tagging and hierarchical tree building.

On the other hand, there are many Chinese word segmentation tools such as Lucene, SCWS, ICTCLAS [6], HTTPCWS, LTP [7], Stanford Word Segmenter [3] and IKAnalyzer. Among them the ICTCLAS system is the most popular one. It has owned several first prizes, but it does not show the advantages when dealing with OBI because the precision rate is only 50% when it was used to realize the auto-segment of OBI.

Since the OBI has its own grammatical features, it is necessary to study the special word segmentation method for OBI. Therefore, this paper aims to study word segmentation of OBI annotations. We considered the word frequency, part of speech and OBI dictionary. The innovation of this paper is using the combination of OBI electronic dictionary and sliding window algorithm to deal with the problem of OBI word segmentation. And we first put forward the concept of computational oracle bone inscriptions.

OBI grammar features

Though OBI is the earliest writing system of China, its many characteristics are extended to the modern Chinese, such as pictographs, knowing, phonetic, self explanatory, transfer, under the guise of such characters, grammar and syntax. Cai Huiying et al. [8] studied the 2703 OBI characters in authoritative Monographs Oracle Bone Inscriptions Dictionary by plane survey, it was found that in the oracle bone inscriptions monosyllabic words dominant, dominant monosemic words, accounting for 77.2% in the total number of monosemic words, far beyond the synonyms; the number of noun is greatly outnumber than the number of verb; proper nouns overwhelming advantage of nouns, and the monosemic words in the proper noun accounted for 81.7% and the polysemy words in the proper noun accounted for 77.5% ; the number of content words is much more than that of empty words, the total number of empty words is 26 and the single word is 26, accounting for 4.5% of the total number of monosemic words, the polysemous words are 20 in word meaning, accounting for 3.3%.

As the first form of Chinese grammar and text system, OBI has many features extending to future generations literature handed down from ancient times. But it has some different characteristics from other ancient writing: 1) special-shaped words; 2) different words have the same font; 3) combination font is a common phenomenon, that is, there are two or three words

merged together, but only occupy one word's position [9]; 4) a high-frequency words accounts for a high proportion of the total accounted for low frequency words and word very low proportion accounted for high proportion of the total vocabulary at both ends of the concentration of [10] in the total quantity of words; 5) three object verb is a unique phenomenon [11]; 6) there are four part of the integrity OBI inscriptions: front inscription, command inscription, divination inscription and verification inscription, but most inscriptions are omitted some parts, common inscriptions retained only the front and command ones [9].

OBI electronic dictionary

We built an OBI electronic dictionary according to above OBI grammatical features because OBI dictionary is the basis of the word segmentation. The structure of the dictionary is designed according to the part of speech and used frequency of the words. It is important to consider the maintainability and scalability when the dictionary is designed. Table 1 shows the dictionary structure.

Table 1. The structure of OBI dictionary

No.	The fields of OBI dictionary		
	Column name	Data type	Declaration
1	Id	char(6)	the number of OBI word
2	Jtz	varchar(20)	simplified Chinese character
3	Ftz	varchar(20)	original complex form
4	Ldz	varchar(20)	redefined word
5	Jgz	varchar(20)	oracle bone character
6	Bh_set	varchar(20)	oracle bone serial number
7	Cl	char(1)	part of speech
8	Cp	int	word frequency
9	Yylb	varchar(10)	semantic category
10	Load	boolean	is the word in memory or not, false is default

Currently the OBI dictionary contains 4881 entries (including variant OBI characters and combination font). There are 4687 single words, 174 two-character-words and 20 three-character words.

OBI word segmentation

The process of our OBI word segmentation method is as follows: first, extract the corresponding inscriptions text from OBI corpus to. Second, finish the rough word segmentation based on dictionary. Third, parse the syntax through bottom-up method. Fourth, handle with the ambiguous words and unlisted words using syntactic rule base, checking the unknown words and recall the valid ones into the OBI dictionary at the same time. Last, optimize the result of word segmentation. The OBI word segmentation flow is shown in Fig. 1.

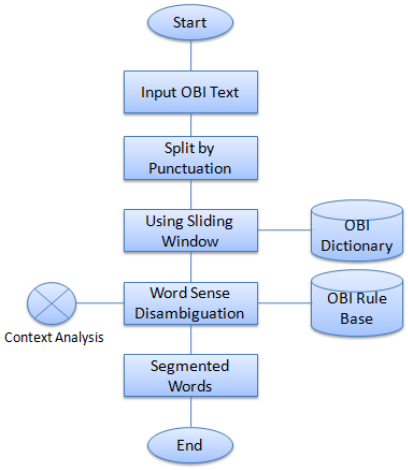

Figure 1. OBI word segmentation flow

Sliding window for obi word segmentation

Sliding window is originally an algorithm in computer network, now it is used here to achieve the OBI word segmentation. We use a variable size window to restrain the OBI inscriptions, according to the size of the window to display the content is a word or not in the OBI dictionary to realize word segmentation. The algorithm is as follows.

Input: OBI inscriptions $O_0O_1O_2...O_{n-1}$ and window length len .

Output: Segmented OBI words.

Algorithm:

for i=0 to n-1

for j=0 to n-len

if substring $O_{i-1}O_i...O_{i+len-2}$ is included in OBI dictionary but any other substring such as $O_iO_{i+1}...O_{i+len-1}$, $O_{i+1}...O_{i+len-1}O_{i+len}$ are not included in the dictionary

then, $O_{i-1}O_i...O_{i+len-2}$ is an OBI word and remove it from the OBI inscription;

i=i+len-1;

j=i+len-1;

elseif neither substring shown in the window is included in OBI dictionary

i++;

j++;

else

do word sense disambiguation.

Support the size of sliding window is 3, the algorithm initialization is shown in Fig. 2.

Figure 2. Algorithm initialization

In fact, the size of sliding window can be changed as n, and $n \in [2, len_{\max}], len_{\max}$ means the maximum word length of OBI dictionary.

Fig. 3 shows the certain condition that not only one substring appears in the OBI dictionary which its length is 3. It needs to be disambiguated.

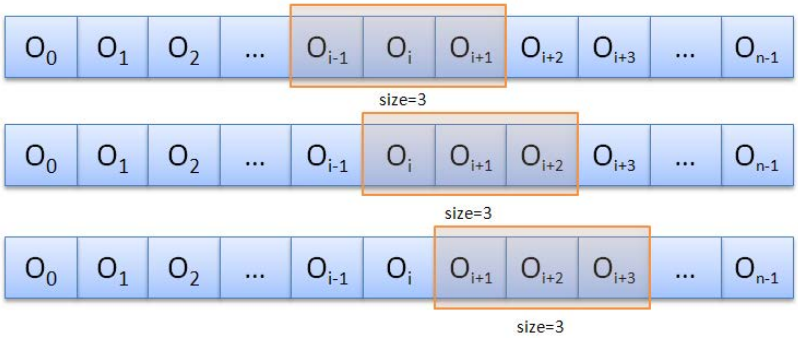

Figure 3. Two or three 3-length words in the OBI dictionary when the window size is 3

Because the most words in OBI dictionary are single words, so we set the window size are 2 and 3. All of the words which not appear in the sliding window at last are automatically segmented as single words.

OBI word sense disambiguation

In OBI many words are variety of parts of speech. For example, a certain word is both a noun and a verb. It probably causes some problems when splitting it. In this case, we should use OBI syntactic rules to deal with the word sense disambiguation.

There are three level OBI syntax forms, including words, phrases and sentences. We can gain words from OBI dictionary automatically, and phrases are composing of words and they are components of the sentences. The following summary gives some common grammar rules discussed in [6]. The structure type of OBI phrase has 11 categories. OBI sentence type can be divided into predicate sentence, non-subject-predicate sentence, elliptical sentence and inversions.

OBI phrase structure as shown in Table 2 where SP means subject, PP means predicate, OP means object, AtP means attribute, AdP means adverbial, CP means complement.

Table 2.OBI phrase categories

No.	OBI phrase categories and description		
	Structure type	Function type	Formal language description
1	subject predicate	single sentence	SP+PP→P1 : DJ
2	verb-object short	verb	PP+PP→P2 : V

3	verb-complement	verb/adj	PP+CP→P3 : V,A
4	modifier-noun	noun	AtP+N→P4 : N
5	adverbial-verb	verb	AdP+N→P5 : V
6	conjunction-redicate	verb	SP+PP+PP→P6 : V
7	concurrent	verb	V+N+V→P7 : V
8	coordinate	verb/noun/adv	V,N,D→P8 : V,N,D
9	appositive	noun	N+N→P9 : N
10	quantitative	noun	M+Q→P10 : N
11	preposition-object	adv	P+N→P11 : D

Similarly, according to OBI's characteristics we can build OBI sentence form base based on the above structures.

Based on the OBI syntactic rules base, we can deal with the case which one word has variety of parts of speech, but it is not enough to solve the problem shown in Fig. 3. We need to analyze the OBI word language environment or context.

OBI context analysis

As shown in Fig. 3, there is not only one result of the word segmentation, which one should be the best one? The context of the word should be considered. We have done the statistical analysis of occurrence frequency of all the characters in the OBI corpus and the co-occurrence statistics with the previous and the latter characters. It can help us decide which word segmentation is the best one by through the co-occurrence probability. Fig. 4 shows the co-occurrence statistical data of the given OBI character.

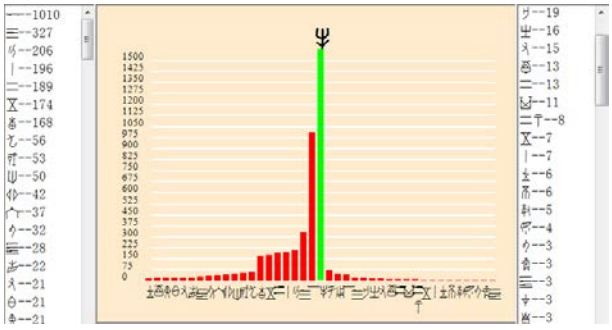

Figure 4. The example of OBI word context analysis

In Fig.4, the center part of the figure shows the context statistical analysis of a certain OBI word, the left part of the figure shows the occurrence frequency of the words appear before the given word, and the right part of the figure shows the frequency of the words after the given word. From the figure we can visually see the occurrence law match with the words before and after the given. In order to solve the problem described in Fig. 3, we can calculate the context analysis situations of the OBI characters appearing in the sliding window and choose the largest one as the final segmentation result.

Experiment and analysis

We used 72151 oracle bone divination text including 6199 words as experimental sample. There are 4881 words in the OBI dictionary. Some common Chinese word segmentation tools and methods are chosen to compare with ours. The precision, recall rates and F-Measure of our word segmentation method comparing with other tools including ICTCLAS, LTP, StanfordSegmenter, HLSegment and PanGu are shown in Table 3.

Table 3. Our word segmentation performance comparing with other tools and methods

Chinese word segmentation tools	Comparison parameter		
	Precision (%)	Recall (%)	F-Measure (%)
ICTCLAS	50.00	47.62	48.72
LTP	37.78	24.60	29.58
StanfordSegmenter	47.14	53.17	49.53
HLSegment	66.67	33.33	44.44
PanGu	59.37	46.03	50.75
Our Method	91.36	94.25	92.78

From Table 3 we can draw a conclusion that our method fits the OBI information processing, and its efficiency is much higher than other methods which oriented modern Chinese

word segmentation. Moreover, when we calculate these parameters the problem of part of speech is not considered. The main reason is that the method based on OBI special dictionary and it follows the syntax rules of OBI and we use context analysis to support OBI sliding window algorithm. Higher recall rate because the vast majority words in the OBI are one-character words. It was also found that word segmentation accuracy rate increase with the augment of OBI Corpus.

Conclusions

Our word segmentation method is based on OBI dictionary and a new algorithm. It is composed of dictionary knowledge and grammar rules. In the future, we will study deeply in speech tagging, syntactic analysis and semantic analysis based on ontology. So the OBI dictionary and syntactic rules base need to be improved and the OBI ontology construction is also an important work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61401242), Grant Henan Youth Core Teacher of University (No. 2014GGJS-106) and the Innovation Talents Support Plan of Science and Technology of Colleges and Universities in Henan Province (No. 15HASTIT023).

References

1. Dong Yuan, Jing Xiong, Yongge Liu (2012) Research on Example-based Machine Translation for Oracle Bone Inscriptions. *New Technology of Library and Information Service*, 28(5), p.p.48-54.
2. Yingying, F., S. Jiqing (2009) Bibliometric Study on Chinese Word Segmentation Papers of China in the Past Five Years. *Journal of Modem Information*, 29 (11), p.p.161-166.
3. Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, Christopher Manning (2005) A conditional random field word segmenter for sighan bakeoff 2005. *Proc. Conf. on the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Korea, p.p. 168-171.
4. Wang, M., R. Voigt, C.D. Manning (2014) Two Knives Cut Better Than One: Chinese Word Segmentation with Dual Decomposition. *Proc. Conf. on the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, USA.
5. Wang, Z., C. Zong, N. Xue (2013) A Lattice-based Framework for Joint Chinese Word Segmentation, POS Tagging and Parsing. *Proc. Conf. on the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, p.p. 884-889.
6. Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, Qun Liu (2003) HHMM-based Chinese lexical analyzer ICTCLAS. *Proc. Conf. on the second SIGHAN workshop on Chinese language processing*, Sapporo, Japan,p.p. 184-187.
7. Wanxiang Che, Zhenghua Li, Ting Liu (2010) Ltp: A chinese language technology platform. *Proc. Conf. on the 23rd International Conference on Computational Linguistics: Demonstrations*, Beijing, China, p.p. 13-16.
8. Huiying Cai, Minghu Jiang, Beixing Deng, Lin Wang (2006) Method combining rule-based and corpus-based approaches for oracle-bone inscription information processing. *Lecture Notes in Computer Science*, 4114, p.p. 736-741.
9. Liu Yi-man (2000) The Characteristics and Main Content of Oracle Bone Inscriptions. *Archives Management*, no.1, p.p. 40-41.
10. Liu Zhi-ji. (2010) On the Two Concentration Features of the Character Frequency of Bone Inscriptions. *Studies in Language and Linguistics*, no.4, p.p. 114-122.
11. ZHENG Ji-e. (2008) A Study of Sacrifice Items in Shang Divinations. *Yindu Journal*, no.2, p.p. 19-22.