PACLIC 20

The 20th Pacific Asia Conference on Language, Information and Computation

Edited by
Tingting He
Maosong Sun
Qunxiu Chen

Proceedings of the Conference

Wuhan, China
1-3 November, 2006

Tsinghua University Press

PACLIC 20

The 20th Pacific Asia Conference on Language, Information and Computation

Edited by
Tingting He
Maosong Sun
Qunxiu Chen

Proceedings of the Conference

Wuhan, China
1-3 November, 2006

Tsinghua University Press
Beijing

图书在版编目（CIP）数据

第 20 届亚太地区语言、信息和计算国际会议论文集=The 20th Pacific Asia Conference on Language, Information and Computation / 何婷婷，孙茂松，陈群秀 主编. —北京：清华大学出版社，2006.10
ISBN 7-302-14060-X

Ⅰ. 第… Ⅱ. ①何… ②孙… ③陈… Ⅲ. ①信息语言－国际学术会议－文集－英文 ②计算语言学－国际学术会议－文集－英文 Ⅳ. H087-53

中国版本图书馆 CIP 数据核字（2006）第 125669 号

出 版 者：清华大学出版社　　地　　址：北京清华大学学研大厦
http://www.tup.com.cn　　邮　　编：100084
社 总 机：0 10-62770175　　客户服务：010-62776969
责任编辑：赵彤伟
封面设计：傅瑞学
印 装 者：北京市清华园胶印 厂
发 行 者：新华书店总店北京 发行所
开　　本：203×280　印张：30.25
版　　次：2006 年 10 月第 1 版　2006 年 10 月第 1 次印刷
书　　号：ISBN 7-302-14060-X/TP・8447
定　　价：200.00 元

Sponsors

National Natural Science Foundation of China
Department of Language Information Processing and Management, Ministry of Education of China
Chinese Information Processing Society of China
Fujitsu Research and Development Center, Co., Ltd.

Organized by

Huazhong Normal University
Chinese Information Processing Society of China

Under the Auspices of

PACLIC Steering Committee

Preface

PACLIC (Pacific Asia Conference on Language, Information and Computation) is a unique inter-disciplinary and multi-lingual conference. It aims to create synergy between theoretical and computational linguists, as well as to provide a forum for exchange and sharing of ideas among scholars of the Pacific Asia region. The strength of the Pacific Asia region lies in its multilingualism and multi-cultural heritage. It is our vision and mission to strengthen this heritage by bridging disciplinary and national boundaries in order to share knowledge and development. We believe that substantial contribution to human beings can be made when synergy is created across the multi-lingual and multi-cultural spectrum. The endurance and longevity of the PACLIC conferences attest to the validity of this vision.

PACLIC20 marks a milestone for two important reasons. In addition to the obvious reason for being the twentieth conference, this is also the first time that PACLIC takes place in China, the most populous and arguably the most linguistically diversified country in this region. China is not only a country whose ancient civilization has long influenced neighbouring cultures, it is also has one of the most vibrantly developing academic and technical sector.

The conference site of Wuhan, a tri-city striding the merging Han and Yangtze rivers, is an auspicious sign for our efforts in pursuit of synergy from multiple sources. At this historical meeting place of different cultures in China, we welcome participants from no less than 10 countries and regions, including Czech Republic, China, Germany, China Hong Kong, Japan, Korea, The Philippines, Singapore, China Taiwan, and Vietnam. We know for sure that our divergent backgrounds will be the sources for our emergent multilingual synergy.

For a successful PACLIC20 that will break new grounds, we thank the thoughtful and tireless organizing work done by Professor Tingting He and her team; as well as the careful and timely paper review by Professor Maosong Sun and his program committee members.

To conclude, we would like to reflect on our intellectual debt to the original PACLICers, although this name was not used then. In Seoul in January 1982, a small group of like-minded linguists from Japan and Korea gathered to discuss in a collegial environment in spite of their linguistic and cultural differences. They lit the visionary fire that is burning brightly for us now. Among this small group, Professor Akira Ikeya (池谷 彰) has remained to be the heart and soul of PACLIC for nearly 25 years. Professor Ikeya has just announced his retirement from PACLIC steering committee, long after his retirement from Toyo Gakuen University. Thank you, Professor Ikeya, for sharing your heart and soul with us and with all future PACLIC conferences.

Chu-Ren Huang and Zhendong Dong
PACLIC20 Conference Co-chairs, and
on behalf of PACLIC Steering Committee Members

Conference Organizers

International Advisory Committee for PACLIC 20

Jun'Ichi Tsujii (University of Tokyo, Japan)
Victor Zue (The Computer Science and Artificial Intelligence Laboratory, MIT, USA)
Chin-Chuan Cheng (Academia Sinica, Taipei, China)
Shiwen Yu (Peking University, China)
Ping Chen (University of Queensland, Australia)
Key-Sun.Choi (Korea Advanced Institute for Science and Technology, Korea)

Steering Committee of PACLIC

Jae-Woong Choe (Korea University, Korea)
Yasunari Harada (Waseda University, Japan)
Chu-Ren Huang (Academia Sinica, Taipei, China)
Akira Ikeya (Tokyo Gakugei University, Japan)
Kim Teng Lua (Chinese and Oriental Languages Information Processing Society, Singapore)
Benjamin T'sou (City University of Hong Kong, China)
Tingting He (Huazhong Normal University, China)

Honorary Conference Chair

Yuming Li (The Department of Language Information Processing and Management, The Ministry of Education, China)

Conference Chairs

Zhendong Dong (Research Center of Computer & Language Information Engineering, CAS, China)
Chu-Ren Huang (Academia Sinica, Taipei, China)

Program Committee Chair

Maosong Sun (Tsinghua University, China)

Program Committee Co-Chairs

Donghong Ji (Institute for Infocomm Research, Singapore)
Qin Lu (Polytechnic University of Hong Kong, China)
Mei-Chun Liu (Chiao Tung University, Taipei, China)
Yongkyoon No (Chungnam National University, Korea)
Yuji Matsumoto (Nara Institute of Science and Technology, Japan)

Program Committee

Akira Ikeya (Tokyo Gakugei University, Japan)
Alex Chengyu Fang (City University of Hong Kong, China)
Alexander Gelbukh (Instituto Politecnico Nacional, Mexico)
Hang Li (Microsoft Research Asia, China)
Tingting He (Huazhong Normal University, China)

Xuanjing Huang (Fudan University, China)
Jian-Yun Nie (Universite de Montreal, Canada)
Jong C. Park (Korea Advanced Institute for Science and Technology, Korea)
Keh-Jiann Chen (Academia Sinica, Taipei, China)
Kiyong Lee (Korea University, Korea)
Kiyotaka Uchimoto (National Institute of Information and Communications Technology, Japan)
Laurent Romary (Laboratoire Loria, CNRS, France)
Shaoming Liu (Corporate Research Group, Fuji Xerox, Co., Ltd., Japan)
Monte George (American National Standards Institute, USA)
Nianwen Xue (University of Colorado, USA)
Qing Ma (Ryukoku University, Japan)
Yan Qu (Clairvoyance Corporation, USA)
Satoshi Tojo (Japan Advanced Institute of Science and Technology, Japan)
Virach Sornlertlamvanich (Thai Computational Linguistics Laboratory, NICT, Thailand)
Houfeng Wang (Institute of Computational Linguistics, Peking University, China)
Xiaojie Wang (Beijing University of Posts and Telecommunications, China)
Yasunari Harada (Waseda University, Japan)
Jun Zhao (Institute of Automation, Chinese Academy of Sciences, China)
Guodong Zhou (SuZhou University, China)

Local Organizing Committee Co-Chairs

Tingting He (Huazhong Normal University, China)
Youqi Cao (Chinese Information Processing Society of China, China)
Qunxiu Chen (Tsinghua University, China)

Invited Reviewers

Qun Liu (Institute of Computing Technology, Chinese Academy of Sciences, China)
Changping Liu (Institute of Automation, Chinese Academy of Sciences, China)
Quan Zhang (Institute of Acoustics, Chinese Academy of Sciences, China)
Jingbo Zhu (Northeastern University, China)
Junfeng Hu (Peking University, China)
Min Zhang (Tsinghua University, China)
Sujian Li (Peking University, China)
Shaoming Liu (Fuji Xerox Co., Ltd., Japan)

Table of Contents

Which Is Essential for Chinese Word Segmentation: Character versus Word

Chang-Ning Huang and Hai Zhao

Microsoft Research Asia,
49, Zhichun Road, Haidian District,
Beijing, China, 100080
{cnhuang, f-hzhao}@msrchina.research.microsoft.com

Abstract. This paper proposes an empirical comparison between word-based method and character-based method for Chinese word segmentation. In three Chinese word segmentation Bakeoffs, character-based method quickly rose as a mainstream technique in this field. We disclose the linguistic background and statistical feature behind this observation. Also, an empirical study between word-based method and character-based method are performed. Our results show that character-based method alone can work well for Chinese word segmentation without additional explicit word information from training corpus.

1 Introduction

Chinese text is written without natural delimiters, so word segmentation is an essential first step in Chinese language processing. In this aspect, Chinese is quite different from English in which sentences of words delimited by white spaces. Though it seems very simple, Chinese word segmentation (CWS) is not a trivial problem. Actually, it has been active area of research in computational linguistics for almost 20 years and has drawn more and more attention in the Chinese language processing community. To accomplish such a task, various technologies are developed [1][2].

In the early work of Chinese word segmentation, word-based method once played the dominant role, in which maximum matching algorithm is the most typical method. Here, the term, word, means those known words are shown in known lexicon or training corpus (also are called in-vocabulary(IV) words.). Explicit known word information was still important learning object even after statistical methods were introduced in CWS [1].

To give a comprehensive comparison of Chinese segmentation on common test corpora, three International Chinese Word Segmentation Bakeoffs were held in 2003, 2005, and 2006[1], and there were 12, 23 and 23 participants, respectively [3], [4], [5]. Four segmentation corpora were presented in each Bakeoff. Thus, twelve corpora are available from Bakeoff 2003, 2005, and 2006. A summary of these corpora is shown in Table 1.

In all of proposed methods, character-based tagging method [6], instead of traditional word-based one, quickly rose in Bakeoff-2005 as a remarkable one with state-of-the-art performance. Especially, two participants, Ng and Tseng, gave the best results

[1] In 2006, the name of the third Bakeoff has been changed into International Chinese Language Processing Bakeoff for the reason that named entity recognition task was added

in almost all tracks [7], [8]. In Bakeoff-2006, all participants whose system performance ranked first in a track at least used character-based method. Researchers turned to character-based method from traditional word-based method only with four years.

The success of Bakeoffs not only gave some public consistent segmentation standards, but also proposed a corpus-based segmentation standard representation, instead of the representation of known word lexicon and segmentation manual before. Thus Chinese word segmentation becomes more like corpus-based machine learning procedure in this sense.

With the supply of common segmentation standards of Bakeoffs, the comparison problem on word-based method and character-based method are still remained. Though most effective Chinese word segmentation techniques are turned to pure character-based methods, some researchers are still insisting that character-based method alone can not be superior to the method that combines both word information and character information [9] [10][11]. In this paper, we will briefly explore the linguistic background of such turnaround in Chinese word segmentation and give an empirical comparison of these methods.

Table 1. Corpora statistics of Bakeoff 2003, 2005 and 2006

Provider	Corpus	Encoding	#Training words	#Test words	OOV rate
Academia Sinica	AS2003	Big5	5.8M	12K	0.022
	AS2005	Big5	5.45M	122K	0.043
	AS2006	Big5	5.45M	91K	0.042
Hong Kong City University	CityU2003	Big5	240K	35K	0.071
	CityU2005	Big5	1.46M	41K	0.074
	CityU2006	Big5	1.64M	220K	0.040
University of Pennsylvania	CTB2003	GB	250K	40K	0.181
	CTB2006	GB	508K	154K	0.088
Microsoft Research Asia	MSRA2005	GB	2.37M	107K	0.026
	MSRA2006	GB	1.26M	100K	0.034
Peking University	PKU2003	GB	1.1M	17K	0.069
	PKU2005	GB	1.1M	104K	0.058

The remainder of the paper is organized as follows. The next section reviews the track of character-based method. We discuss the linguistic background of character-based features (especially for unigram feature) in Section 3. We evaluate unigram feature through CWS performance comparison in Section 4. In Section 5, the experimental

results between word-based method and character-based method are demonstrated. We summarize our contribution in Section 6.

2 The Track of Character-based Method

Character-based tagging method is a classification technique for Chinese characters according to their positions occurring in Chinese words. This method was first conducted in [12], two classifiers were combined to perform Chinese word segmentation. First, a maximum entropy model was used to segment the text, and then an error driven transformation model was used to correct the word boundaries. This method was continuously improved in [6] and [13], where a unified maximum entropy model was used to perform character-based tagging task.

As mentioned above, two top participants, Tseng and Low, won the most outstanding success in Bakeoff-2005 with the similar character-based tagging method, though the former used conditional random field model while the latter still used maximum entropy model.

In Bakeoff-2006, all participants whose system performance ranked first in a track at least used character-based method. There are five participants ranked the first in one track at least [14][15][16][17][18], in which two participants used conditional random field, and the other three used maximum entropy as learning model. Especially, four participants directly or indirectly used the technique in [7].

3 Features of Character Classification for CWS

CWS is the primary processing in Chinese language processing. Thus it is difficult or even impossible to use derivative features like other Chinese language processing tasks. The basic features that we can use are characters themselves.

We perform a position frequency statistics of Chinese characters in MSRA2005 training corpus. All characters appearing in this corpus are counted. Six positions are distinguished, which are represented by a 6-tag set including B, E, S, B_2, B_3, and M [14]. Tag B and E stand for the first and the last position in a multi-character word, respectively. S stands up a single-character word. B_2 and B_3 stand for the second and the third position in a multi-character word, whose length is larger than two-character or three-character. M stands for the fourth or more rear position in a multi-character word, whose length is larger than four-character.

Let $T = \{B, E, S, B_2, B_3, M\}$, we calculate the *productivity*, $P_{C_i}(t_j)$, of each position of each character C_i:

$$P_{C_i}(t_j) = \frac{count(C_i, t_j)}{\sum_{t_j \in T} count(C_i, t_j)} \quad (1)$$

We count those characters whose productivity is larger than 0.5 for a certain tag. The results are shown in Table 2. There are 5,147 different characters in MSRA2005 training corpus. Our statistics shows that most characters, 76.16% of all, trend to have a stable position in the word. This is important for a character-based tagging method. However,

there are still 1,227 characters without dominant tag. We regard these characters as free ones. The fact that no special positions are dominant for a character means that this character can occur in every possible positions in a word. That is, this character is free for word formation. In our threshold of productivity 0.5, 1/4 characters (precisely, 23.84%) in one of real corpora, MSRA2005, are free ones.

Table 2. The distribution of numbers of characters in each position

Tag	B	B_2	B_3	M	E	S	Total
Number of characters	1634	156	27	33	1438	632	3920
Percent(%)	31.74	3.03	0.52	0.64	27.94	12.28	76.16

We list ten top frequent characters and their tag distributions in Table 3 according to MSRA2005 training corpus.

Table 3. Top frequent characters and their tag distributions

Characters	Frequency	B	E	S	B_2	B_3	M
的	129132	0.001169	0.010338	**0.987679**	0.000519	0.000163	0.000132
一	40189	**0.540023**	0.058648	0.285650	0.086889	0.019408	0.009381
国	40091	0.310070	**0.468609**	0.020828	0.151206	0.024968	0.024320
在	32594	0.024821	0.099742	**0.869485**	0.003712	0.002178	0.000061
中	29762	**0.490558**	0.093609	0.315570	0.032424	0.032323	0.035515
了	29305	0.026480	0.052346	**0.919980**	0.000478	0.000682	0.000034
是	28020	0.015703	0.338829	**0.641113**	0.001642	0.002712	0.000000
人	27260	**0.355026**	0.304952	0.228833	0.023844	0.063243	0.024101
和	26328	0.047820	0.008356	**0.922440**	0.007710	0.001785	0.011888
有	26196	0.268133	0.313597	**0.376661**	0.018934	0.008207	0.014468

To demonstrate the distribution of characters with different productivity as single-character word, we count different types of characters in certain range, A bar figure is shown in 1. This figure further shows that most characters trend to be components of multi-character words, instead of single-character words. Especially, more than half of characters nearly never be a single-character word. This is another obvious statistical characteristic for word formation from character combination.

Another convenience for character-based method is that it can be more easily to handle out-of-vocabulary (OOV) words. As well known, the set of all Chinese characters is almost a closed set. 2,500 Chinese characters can cover 97.97% text that one can meet in his life, while 3,500 characters can cover more than 99.48% text[2]. We see that the OOV rate of word for MSRA2005 corpus is 2.6%, while the OOV rate of character is only 0.42% (12 OOV characters versus 2,837 characters in MSRA2005 test corpus). We see that the former is much larger than the latter. In addition, six of these OOV characters appear only once.

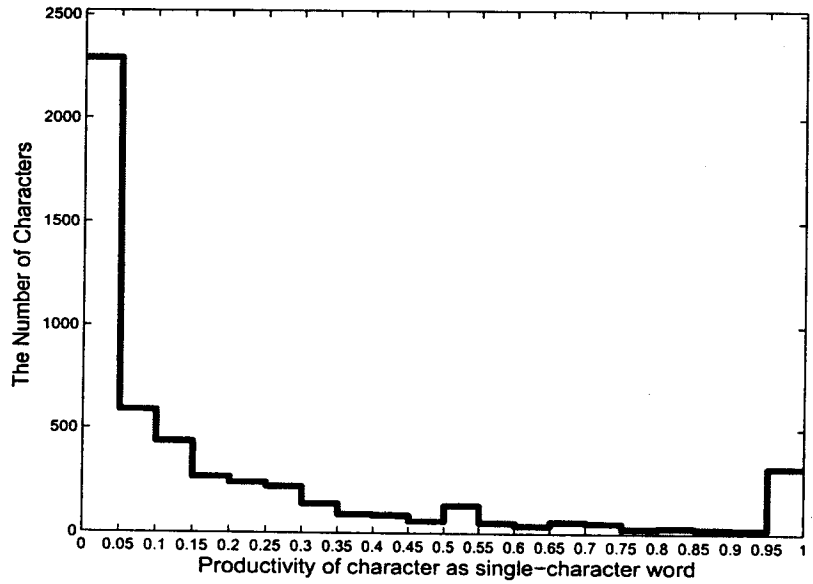

Fig. 1. Character distribution with different productivity as single-character word. All counting are performed when $P_{C_i}(S) \geq j * 0.05$ and $P_{C_i}(S) < 0.05 + j * 0.05$, where $j = 0, 1, ..., 20$.

The productivity of character is the concept of linguistics, while it is just the learning goal as the unigram feature for a sequence learning model. If the context is free or absent, then what a character itself should be a word alone is determined by the productivity of position "S", and what it should be the begging of a word is determined by the productivity of position "B", and so on. Note that segmentation is an operation to determine separation of sequence or not at the current character. The usefulness of productivity of character, or namely unigram feature in learning model, is obvious.

[2] *The introduction to modern Chinese character list in common use (*现代汉语常用字表说明*)*, published by the State Language Affairs Commission and the State Education Commission on January 26th, 1988.

Since most characters trend to be in stable position in word formation, it will be efficient for a character-based classification technique for CWS. One remained challenging thing is the task to determine those characters that can freely appear in each position of words without favoritism, whose percent is 23.84% in all kinds of characters. This leads to more strict context to perform the task to determine the classification of these free characters.

In a character sequence, the straightforward way to represent context is using adjacent characters. Actually, this means that more n-gram features are used. We explain this case in a real sentence, "葡萄是红的(The grape is red.)". The final segmentation result will be "葡萄/是/红/的". In a bigram sense, the reason of such segmentation is bigram probability of "葡-萄" to be a word is much higher than any other bigram probabilities of "萄-是", "是-红" and "红-的". Thus, "葡萄" is finally recognized as a word.

In most Chinese word segmentation systems, all possible n-gram features in a certain character-window of sequence are often used. The difference among them is the length of this character-window. Three-character window and five-character window centered by the current character are mostly adopted in existing work until now.

4 How Unigram Feature Affect the CWS Performance

We adopt the character-based CWS system that was described in [14] in this paper. The learning model is conditional random field [20], and tag set is 6-tag set as mentioned above. However, all none n-gram features in [14] are removed, and feature template list is shown in Table 4. The reason is to conform to the constraints of closed test in Bakeoff, and all features that are beyond provided training corpus are not allowed. All comparisons below will be performed in closed test settings for a consistent circumstance.

Table 4. Feature templates

Code	Type	Feature	Function
a	Unigram	$C_n, n = -1, 0, 1$	The previous (current, next) character
b	Bigram	$C_nC_{n+1}, n = -1, 0$	The previous (next) character and current character
		$C_{-1}C_1$	The previous character and next character

We explain these selected features from a real sentence, "我们在北京" (We are in Beijing). If the current character is "在", then all active features will be "们","在", "北", "们在", "在北", and "们北".

We give a performance comparison among different types of n-gram features and forward maximum matching (FMM) algorithm in MSRA2006 corpus. As for FMM algorithm, we use two dictionaries, one is extracted from training corpus, the other is