北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

Natural Annotation Research in Large-Scale Corpora with a Focus on Chinese Word Segmentation

RAO Gaoqi, XIU Chi, XUN Endong†

College of Information Science, Beijing Language and Culture University, Beijing 100083;
† Corresponding author, E-mail: edxun@blcu.edu.cn

Abstract The distribution and meaning of natural annotations on large datasets are discussed. The proposed research on word extraction shows the positive potential of both implicit and explicit natural annotation in word segmentation. Experiments on word extraction indicates that the implicit natural annotation derived from language laws and patterns are more powerful in splitting character strings in raw corpora.
Key words natural annotation; Chinese word segmentation; word extraction; large-scale corpora

语料库自然标注信息与中文分词应用研究

饶高琦　修驰　荀恩东†

北京语言大学信息科学学院, 北京 100083; † 通信作者, E-mail: edxun@blcu.edu.cn

摘要 以中文分词为应用目标, 将大规模语料库上存在的自然标注信息分为显性标注信息与隐性标注信息, 分别考察了它们的分布和对大数据集上语言计算的影响。结果表明, 两者都直接或间接地表达了作者对语言的分割意志, 因而对分词具有积极的影响。通过词语抽取测试, 发现在缺乏丰富显性标注信息的文本中, 来自语言固有规律的自然标注信息对字符串有着强大的分割性能。
关键词 自然标注信息; 中文分词; 词语抽取; 大规模语料库
中图分类号 TP391

Manually annotated corpora play a significant role in the research and development of Chinese language processing. But their scales are limited by huge cost of manual annotation. Otherwise, Zipf Law decides that data sparseness widely exists in almost all aspects of natural language processing (NLP), which leads to a serious contradiction with the limited scale of manually annotated corpora. Besides the problem of cost, in large-scale corpora, the efficiency and consistency of linguist knowledge formalization form severe challenges. Their quality control in teamwork even adds insult to injury. As the fundamental work of Chinese language processing, Chinese word segmentation achieved great progress but also faced challenges, since its birth. For the methods using machine learning in Chinese word segmentation, exponentially increase of training corpora could only lead to leaner growth of F-figure in segmentation[1].

As an agreement on the exact definition of what a word is remains hard to reach, no automatic and manual method could approach a complete level.

Facing all these challenges, turning to natural annotation in corpora is a natural choice. The information natural annotation offers comes from its natural existence in corpora, instead of exogenous input from experts. This forms an attractive cost

国家自然科学基金(60973062, 61170162)和中央高校基本科研业务费专项资金(2012–jbyz–001)资助
收稿日期: 2012–05–31; 修回日期: 2012–08–28; 网络出版时间: 2012-10-26 17:55
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1755.020.html

advantage. The resources of natural annotation are authors' will and the regular patterns of authors' using language, which partially achieves the challenge of formalization of linguist knowledge (or we could say the natural annotation just comes from linguist phenomena itself).

In web resources, natural annotation is synonym for User Generated Data (UGD)[2], the mining of which has obtained surprising results in many applications[3–6]. In NLP, things are quite similar. Since the sole reason of the written form of language is to represent the spoken[7], the existence of written form could be seen as the largest scale user annotation work to a language. Also the written texts itself are generated by authors/language users. It is not a surprise that, similar research in large-scale corpora is in need and will surely be useful as UDG in web resources.

Obviously, different forms of natural annotation contribute to various applications. In Chinese word segmentation, some methods are utilizing some forms of natural annotation, especially punctuations, as a feature in statistical machine learning models[8–11].

1 Several Word Segmentation Oriented Natural Annotations

Chinese word segmentation could be considered as a problem of word boundary recognition. Our focus on the word boundaries entails the autonomy of characters or symbols.

Note that, all characters and symbols in a language has the possibility to be a word boundary, that is to say, they are more or less carrying boundary information. We view all the characters and symbols as boundary information carrier (BIC). Once a BIC appears in a string, the string it belongs to could be divided into three substrings (may not be correct), as following:

$$C_1C_2\ldots C_i\mathrm{BIC}_{i+2}C_{i+3}\ldots C_n.$$

If the context of a BIC is assumed to be random and independent, the autonomy of a BIC with strong segmentation ability means that it never or rarely associates with other characters or symbols in this language and its division of strings are mostly correct. In this paper, the BICs of this property are defined as "segmentation indicator" (SI).

IWP (independent word probability) is used to describe the autonomy of BICs[12]. Naturally the top ones in the IWP rank are given priority to be added into SI series for segmentation.

$$\mathrm{IWP}=\frac{N(\mathrm{Word}(c))}{N(c)},$$

c denotes a character and Word(c) denotes that c appears as an independent word. $N(\cdot)$ is its count in corpus.

Definitely a higher IWP means a lower association of a BIC. In corpus of 1993–2003 People's Daily, the top BICs are shown in the IWP rank in Table 1.

Note that, the BIC frequency is not considered in the calculation of IWP. Some rare or ancient characters can get very high IWP value, due to the monosyllables of ancient Chinese[13]. As Table 1 shows, the characters with IWP value of 1 like "鬆", "滹" and "仂"① could be perfect SIs, but their low

Table 1 Top 691 BICs in IWP rank of 1993-2003 People's Daily corpus

Rank	BIC	IWP	Rank	BIC	IWP
1	^P (hard return)	1	1	%	1
1	,	1	1	鬆	1
1	。	1	1	滹	1
1	!	1	1	螅	1
1	?	1	1	仂	1
1	:	1	1	匾	1
1	“	1	1	楮	1
1	”	1	1	谝	1
1	（	1	1	稃	1
1	）	1	1	攉	1
1	——	1	…	…	…
1	【	1	689	哩	0.998296
1	】	1	690	铱	0.995984
1	《	1	691	的	0.988948
1	》	1			

① Means "coiffure on top of the head" ,"descriptive of floodwater or torrential", "surplus", respectively.

frequencies (respectively 6, 5 and 7 times in whole corpus) effect little in future application. Therefore IWP is developed to IWS (independent word strength) by adding the frequency into the formula of IWP.

$$\mathrm{IWS}(c) = \mathrm{IWP}(c)P^2(c) = \mathrm{IWP}(c)\frac{\mathrm{freq}(c)^2}{N^2}.$$

freq(·) indicates the frequency of a character or a symbol in corpus and N is the count of all characters and symbols of corpus.

In Table 2, punctuations have perfect IWS value and undoubtedly they are ideal SIs. The Chinese characters with high IWS value can also function as SIs. Take "的" as an example. Except for very limited words associated with other characters like "的士" and "的确", "的" always appears independently as an auxiliary word or pronoun. Similarly characters like "和", "在", "是"[①] etc also have high IWS value. We define these characters, which can be used as SIs, intensive independent word (IIW).

Great assimilation of Chinese language leads to rare appearance of the compound words mixed with symbols from foreign language like Latin letters and Arabic numerals. Words like "Java 语言", "K 联赛", "211 工程"[②] take only 0.49% in the Golden Standard Lexicon (GSL, manual segmentation by Peking University) of 1998 People's Daily. Therefore Latin letters and Arabic numerals are also very good SIs.

Punctuations and paragraphing are grouped as explicit SIs, while Arabic numerals, Latin letters and intensive independent words as implicit SI. Because the former exists to segment speech and the latter has its own semantic purpose other than segmentation. In Section 2, we'll conduct experiments respectively to measure their segmentation ability.

Table 2 Top 13 in IWS rank of 1993–2003 People's Daily corpus

BIC	Rank	IWS	BIC	Rank	IWS
的	1	0.033582	是	8	0.003267
，	2	0.031799	“”	9	0.003058
。	3	0.013281	！	10	0.002924
、	4	0.008178	一	11	0.002400
了	5	0.006212	月	12	0.002008
和	6	0.006123	个	13	0.001910
在	7	0.005626			

2 Experiments

Two experiments are conducted to describe the segmentation ability of explicit SIs (mainly punctuations) and implicit SIs (mainly intensive independent words).

2.1 Evaluation

The conformity between vocabulary words and the substrings left by SIs segmentation is a clue to observe segmentation ability of various SIs. Note that, if great amount of substrings are vocabulary words, then this consistency could be described by recall rate of the GSL, the formula is as following:

$$\mathrm{Recall} = \frac{N(\mathrm{GSL} \cap \mathrm{SubSet})}{N(\mathrm{GSL})},$$

$N(\cdot)$ is the count and GSL denotes word set of Golden Standard Lexicon, SubSet is the substring set formed by SIs' segmentation.

If the distances between SIs and a particular kind of words are binarized as "two sides adjacent or not":

$\mathrm{SI}_1 C_1 \ldots C_n \mathrm{SI}_2$, two sides adjacent,

$\left.\begin{array}{l} \mathrm{SI}\ C_1 \ldots C_n \\ C_1 \ldots C_n \mathrm{SI} \end{array}\right\}$ one side adjacent.

This recall rate can be viewed as an average description to demonstrate the relative positions of various words in a sentence. This will be further discussed in Section 2.4.

Since a whole corpus could be seen as a string, we could get a substring set when the SIs discussed in Section 1 are utilized to split this string, as following:

$$C_1 C_2 \ldots C_i \mathrm{SI}_1 C_{i+2} C_{i+3} \ldots C_n \mathrm{SI}_2 C_{n+2} C_{n+3} \ldots C_m.$$

C_{i+1} and C_{n+1} are supposed to be two SIs, SI_1 and SI_2, in the string $C_1 C_2 \ldots C_m$. Split by SI_1 and SI_2, the substring set formed after segmentation is: $C_1 C_2 \ldots C_i$, SI_1, $C_{i+2} C_{i+3} \ldots C_n$, SI_2, $C_{n+2} C_{n+3} \ldots C_m$. Overlapped elements should be removed, if there's any. The

① Means "and" ,"at/on/in", "be", respectively.

② Means "Java Programming Language" ,"Korean Football League", "211 Program", respectively.

frequencies of each substring are recorded before removing overlapped ones. We name this frequency "SI Adjunction Frequency", which is different from the real frequency in corpus. Here we assume that SI contains only one character or symbol, though in fact it could contain multi-character.

In order to measure the string segmentation ability of various SIs, the recall rate is calculated by substring set from corpus with the GSL.

2.2 GSL agreement of punctuation, Latin letters and Arabic numerals

Fig. 1(a) demonstrates recall rate changes with the size of corpus with punctuations as SI, and Fig. 1(b) with SI series of punctuations, Latin letter, and Arabic numeral. Golden Standard is the lexicon of People's Daily in 1998 (annotated by PKU standard[14]) for both experiments.

The substring sets for both experiments are formed by People's Daily (year 1998 included). The curve named "total" is the recall rate of the whole vocabulary. "name", "place" and "org" are respectively the recall rate of person names, place names and organization names (Person names, place names and organization names take 26.26%, 10.0%, 23.9% in the GSL, respectively). As demonstrated, the recall rate of the whole corpus outperforms the others.

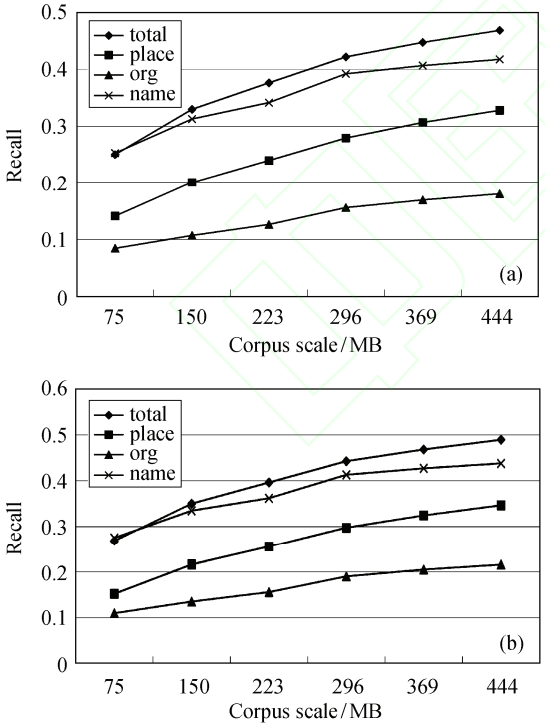

Fig. 1 Punctuations as SI (a) and punctuations, Latin letters, Arabic numerals as SI (b)

Comparing curves in Fig. 1(a) and Fig.1(b), quite limited affect is found of Arabic numerals and Latin letters in the increasing of recall rate (about 0.5%), which conforms to our prior knowledge about Chinese language: punctuations take 9% in all symbols and characters of People's Daily, while Latin letters 2%, Arabic numerals only 1%①, which can explain the little difference brought by Latin letters and Arabic numerals.

Since the source corpus of Golden Standard (People's Daily in year 1998) is included in test, another test without this part of corpus is conducted. In Fig. 2 and Table 3, 1993–2002 People's Daily corpora are used except 1998. However, little change is found compared with Fig. 1(b). The recall rate of the whole vocabulary is still the best, reaching 46.25% using 444 MB corpus.

2.3 GSL agreement with intensive independent word

"的" is the Chinese character with highest IWS value according to the rank in Table 2 and it is the first intensive independent word added into SI series, which means whenever "的" is found in corpus, the string it belongs to will be split into two by it. Table 4 describes the difference this character brings as the only added SI in corpus 1998 People's Daily.

In modern Chinese, "的" is neither used in person names, place names or organization names, nor forms a name itself. Therefore the increase in Table 4 can be thought all caused by the split strings whose context is "的", as following:

的 $C_1C_2C_3\ldots C_iP/N/L$,
的 $C_1C_2C_3\ldots C_i$ 的,
$P/N/L$ $C_1C_2C_3\ldots C_i$ 的。

$C_1C_2\ldots C_i$ is a Chinese string, $P/N/L$ means punctuations, Arabic numerals and Latin letters (the

① A Latin letters string or Arabic numerals string is counted as one time in frequency calculation. For example, in the sentence "和平路 162 号的 H 先生家做的 Nayota 冰激凌很好吃", frequency of Latin letters is 2 and Arabic numerals 1.

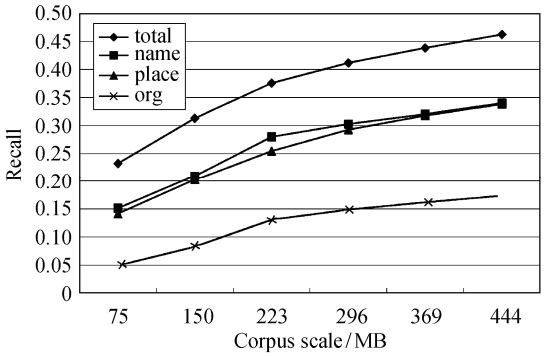

Fig. 2 Punctuation, Arabic numerals and Latin letters in enlarged corpus①

Table 3 Punctuation, Arabic numerals and Latin letters in enlarged corpus

Scale/MB	total	name	place	org
75	0.2324	0.1513	0.1415	0.0493
150	0.3131	0.2094	0.2024	0.0828
223	0.3756	0.2800	0.2545	0.1307
296	0.4117	0.3029	0.2928	0.1486
369	0.4384	0.3204	0.3180	0.1617
444	0.4625	0.3400	0.3381	0.1735

Table 4 Difference brought by "的" as SI

Item	No "的" added/%	Added "的" as SI/%	Increase
Total	25.27	34.22	8.95
Person Names	13.11	15.12	2.01
Place Names	16.93	24.62	7.69
Organizations	14.73	20.51	5.78

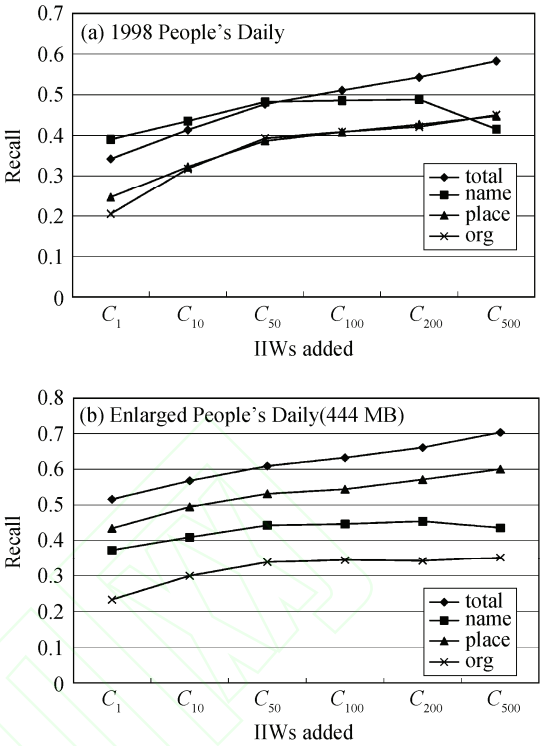

Fig. 3 Intensive independent words added into SI

Table 5 Intensive independent words added into SI

Item	C_1	C_{10}	C_{50}	C_{100}	C_{200}	C_{500}
Total	0.3421	0.4137	0.4772	0.5111	0.5434	**0.583**
Name	0.3905	0.4305	0.4829	0.4862	**0.4887**	0.4158
Place	0.2462	0.3223	0.3869	0.4083	0.4268	**0.4477**
Org	0.2051	0.3178	0.3931	0.4089	0.4215	**0.4505**

SIs already in use). The added Intensive Independent Words in SI series make more substrings split from longer ones. Adding IIWs according to the IWS rank, the recall rate changes as Fig. 3 (a) and Table 5 (experimental corpus is 1998 People's Daily, 12 months in use). C_n indicates the amount of the added intensive independent words.

Significant increase is brought by intensive independent words: almost 25 percentages in recall rate of whole corpus. Decrease from person names is observed when top 500 IWS characters are added into SI series, compared to top 200. Because some new added SI characters are used in person names, leading

Table 6 Intensive independent words as SI in enlarged corpus as in Fig. 3 (b)

Item	C_1	C_{10}	C_{50}	C_{100}	C_{200}	C_{500}
Total	0.515812	0.567945	0.609334	0.632321	0.660913	**0.703355**
name	0.3728	0.4092	0.443	0.4467	**0.4542**	0.4359
place	0.434468	0.494546	0.531382	0.544099	0.571233	**0.600394**
org	0.232673	0.299207	0.337957	0.344085	0.341806	**0.351569**

① The corpus of 1998 People's Daily is 75MB, so here we choose 75MB as an unit. Year boundaries in enlarged corpus are broken, while month boundaries are not.

to more wrongly-split substrings. The curves with similar trend are indicated in Fig. 3 (b) and Table 6, but in limited increase (People's Daily 1998 is not included in 444 MB corpus in test shown by Fig. 3 (b)).

The curves with increasing trend are the result of autonomy of intensive independent words and show the separating capacity of SIs in a language. Our prior knowledge of character choice in Chinese named entities also conforms to the phenomenon shown in Fig. 3: faster increase along with IIW being added indicate that Chinese characters with stronger association are gaining priority in language use, which is resulted by the polysyllabic trend of modern Chinese.

2.4 What is more about the curves: segment distance and distribution feature

The recall rate shown in Fig. 2 and Fig. 3 could be explained in form of Segment Distance, defined as follow. $word_1...word_N$ consist a word set like person names or place names, etc. Segment Distance describes the average distance of a word set (WS) to SIs.

$$\begin{cases} \mathrm{Dis}(\mathrm{word}_i)=0; & \text{if } \mathrm{word}_i \text{ is correctly split,} \\ \mathrm{Dis}(\mathrm{word}_i)=1; & \text{if not,} \end{cases}$$

$$\mathrm{SegmentDis}(\mathrm{WS})=\frac{1}{N}\sum_{i=1}^{N}\mathrm{Dis}(\mathrm{word}_i).$$

In Fig. 3 (a), organization names gain the greatest increase in three kinds of named entities, indicating their different context distribution in corpus. Their SI adjunction frequency has a remarkable rise with more added intensive independent words. This could be thought that organization names have larger segment distance to the SIs of high IWS value. Thus organization names are more likely to hide in the middle of a lot of low IWS value symbols. That is to say, their context is more random and independent than other named entities. Since the high IWS value characters are more likely to be prepositions or auxiliaries, this phenomenon shows the more free syntactical positions of organization names in a sentence. A distribution feature of these named entities is demonstrated by a computational method in the way.

In Fig. 3 (b), recall rates of all 4 items share a similar increasing trend and scale, but different start points, indicating different Segment Distances to punctuations, Latin letters and Arabic numerals. About 60% of words in Golden Standard Lexicon belong to those 3 kinds of name entities. That means the other "normal words" has much smaller Segment Distance to SIs. It requires more deeply research in classifying these words, in order to discover their ability of consistency splitting strings.

3 Rethinking Natural Annotations

Natural annotations, especially the Segment Indicators (SIs), come from authors' segment will and regular patterns of the language in use. Their function on words extraction shows the potential influence of natural annotation in word segmentation.

Not to Process, but to Enlighten: Does "raw corpus" exist in real sense, since rich information has been found from natural annotations in the corpora once thought unprocessed and raw? Instead of information input by annotators and exogenous tagging sets, natural annotation detection requires to "enlighten" the already existed data and let the corpora "say" by themselves.

Jungle or Desert: The boundary between natural and manual annotation is whether they are given before the corpora collected for its original purpose. Take Wikipedia as an example. The taxonomy, hyperlinks, comments and edition records for each entry are mostly from users instead of the original page editor, which could be considered as the result of authors' collective intelligence. Those annotations play a similar role with punctuations in plain texts. Therefore, they are also likely to be grouped in explicit annotations.

Intensive explicit annotations shorten the strings which are purely consisted by Chinese characters, then the effect of implicit annotations of language laws is relative limited. However, for another extreme case, in the frequently-used "raw" corpora (like People's

Daily), in which the punctuations are almost the only embodiment of the authors' segment will. The regular patterns of languages contribute more natural annotations. This is shown in the experiments in Section 2: Intensive Independent Words are representations of the regular patterns of languages. Therefore different kinds of corpora require various detections for implicit or explicit annotations in future work.

Let corpora "say", and let authors "say". This is much better than a round trip back to the pure rule-based rationalism, since the research in NLP has swung too far[15]. Because of the urgent requirement of information from natural annotations, more researches about their natures, features, and usages will have a promising future. Enlightening large-scale corpora could undoubtedly offer new chances and possibilities to both theoretical computational linguistic research and NLP engineering.

References

[1] Zhao H, Song Y, Kit C. How large a corpus do we need: statistical method vs. rule-based method // LREC-2010. Istanbul, 2010: 1672–1677

[2] Sun Maosong. Natural language processing based on naturally annotated web resources. Journal of Chinese Information Processing, 2011, 25(6): 26–32

[3] Wu Fei, Weld D S. Open information extraction using Wikipedia // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010). Uppsala, 2010: 118–127

[4] Qu honghua, Liu Yang. Interactive group suggesting for Twitter // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011). Portland, 2011: 519–523

[5] Si Xiance, Liu Zhiyuan, Sun Maosong. Modeling social annotations via latent reason identification. IEEE Intellingent Systems, 2010, 25(6): 42–49

[6] Kamvar S D, Harris J. We feel fine and searching the emotional web // Proceeding s of the 4th ACM International Conference on Web Search and Data Mining (WSDM-2011). Hong Kong, 2011: 117–126

[7] De Saussure F. Course in general linguistics. Beijing: Foreign Language Teaching and Research Press, 2001: 24

[8] Li Zhongguo, Sun Maosong. Punctuation as implicit annotations for Chinese word segmentation. Computational Linguist, 2009, 35(4): 505–512

[9] Sun Weiwei, Xu Jia. Enhancing Chinese word segmentation using unlabeled data // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-2011). Edinburgh, 2011: 970–979

[10] Zhao Hai, Kit C. An empirical comparison of goodness measures for unsupervised Chinese word segmentation with a unified framework // Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJNLP-2008). Hyderabad, 2008: 9–16

[11] Zhao Hai, Kit C. Integrating unsupervised and supervised word segmentation: the roe of goodness measures // Information Sciences 181. Hong Kong, 2011: 163–183

[12] Wu A, Jiang Zixin. Statistically-enhanced new word identification in a rule-based Chinese system // Proc of the 2nd ACL Chinese Processing Workshop. Hong Kong, 2000: 41–66

[13] Huang Borong, Liao Xudong. Modern Chinese (Volume I). Beijing: High Education Press, 2002: 252

[14] Yu S, Duan H, Zhu X, et al. Specification for corpus processing at Peking University: word segmentation, POS tagging and phonetic notation. Journal of Chinese Language and Computing, 2003, 13 (2): 121–158

[15] Church K. A pendulum swung too far. Linguistic issues in language technology (LiLT), 2011, 6(5): 1–27