How Large a Corpus Do We Need: Statistical Method Versus Rule-based Method

Hai Zhao(赵海)†* Yan Song(宋彦)* Chunyu Kit(揭春雨)*

†MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems,
Department of Computer Science and Engineering, Shanghai Jiao Tong University,
#800, Dongchuan Road, Minhang District, Shanghai, China, 200240
*Department of Chinese, Translation and Linguistics, City University of Hong Kong,
83 Tat Chee Ave., Kowloon, Hong Kong, China
Email: zhaohai@cs.sjtu.edu.cn, yansong@cityu.edu.hk, ctckit@cityu.edu.hk

Abstract

We investigate the impact of input data scale in corpus-based learning using a study style of Zipf's law. In our research, Chinese word segmentation is chosen as the study case and a series of experiments are specially conducted for it, in which two types of segmentation techniques, statistical learning and rule-based methods, are examined. The empirical results show that a linear performance improvement in statistical learning requires an exponential increasing of training corpus size at least. As for the rule-based method, an approximate negative inverse relationship between the performance and the size of the input lexicon can be observed.

1. Introduction

Zipf' s law which was discovered empirically by (Zipf, 1949) is originally about word token distribution in an English corpus. A general version of the law states the probability of w (it could be word token or something else) as,

$$\hat{p}(w) \sim r(w)^{-a} \tag{1}$$

where $r(w)$ is the rank of w in frequency order, and a is a constant. This empirical result implies that serious data sparseness could occur in computational process for natural languages, and meanwhile reveals a relationship between the word token proportion and their coverage in a corpus. In this paper, we consider such a case, if corpus enlargement is the only way to overcome such data sparseness and the learning method is fixed, then what a performance can we expect to obtain for a specific corpus size? Although such an investigation upon the topic was also included in existing works such as (Banko and Brill, 2001), this study still brings some novelties. First of all, we analyze not only machine learning but also rule-based method. Thus a comparison between them can be made, while only machine learning techniques were concerned with in existing works. Secondly, we choose Chinese word segmentation that is formulated as a sequence labeling task as the study case, this is also slightly different from those classification tasks in previous works. Thirdly, our study is based on more strict experimental settings than the previous ones. This will let our conclusion more reliable and make a quantifiable estimation on the impact of corpus magnitude possible for the first time.

2. Experimental Settings

2.1. The Data

The reason to choose Chinese word segmentation for this study is four-fold. Firstly, word segmentation is a simple enough language processing task. It may be easily modeled as a sequence labeling task by using some popular (statistical) machine learning tools such as Conditional Random Fields (CRFs). Secondly, a rule-based method, maximal matching algorithm, is available for this task, which permits a comparison to statistical learning method for supporting our topic. Thirdly, multiple standard large-scale segmentation corpora that are essentially required by this study have been available since Bakeoff-2003[1]. Fourthly, Chinese word segmentation is a word-focus task, while word is usually just the focus in a typical study of Zipf's law.

Table 1: Corpus size in number of characters

Corpus	AS	CityU	UPUC	MSRA
Training (M)	8.44	2.71	0.83	2.17
Test (K)	146.3	364.5	256.5	172.6

Our experiments are carried out in three largest corpora[2] from Bakeoff-2006[3] (Levow, 2006). Corpus size in the number of characters is in Table 1, where each column means a corpus or a segmentation convention. Among four Bakeoffs that have been held, only in Bakeoff-2006, both training corpora and test corpora are large enough and thus suitable for this study.

For the evaluation criterion, word segmentation performance is usually measured by F-score,

$$F = \frac{2RP}{R+P} \tag{2}$$

where the recall R and precision P are respectively the proportions of the correctly segmented words to all words in the gold-standard segmentation and a segmenter's output.

[1]First Chinese word segmentation Bakeoff, the shared task held by SIGHAN, http://www.sighan.org/bakeoff2003.

[2]UPUC corpus is excluded as the size of its training set is not large enough to support our investigation.

[3]Third International Chinese Language Processing Bakeoff, http://www.sighan.org/bakeoff2006.

2.2. The Method

Both statistical machine learning and rule-based methods are considered for the segmentation task in this study. As for the former, existing work shows that Chinese word segmentation can be effectively formulated as character tagging task by using maximum entropy or CRFs (Xue, 2003; Peng et al., 2004; Song et al., 2006). As the state-of-the-art results in recent two Bakeoffs were given by (Zhao et al., 2006a; Zhao et al., 2006b; Zhao and Kit, 2008b; Zhao and Kit, 2008a), it is shown that the CRFs learning achieves a better segmentation performance with a 6-tag set than any other tag set. We opt for using this tag set that represents character position in a word and the corresponding six n-gram feature templates for our experiments.[4] The six tags are B, B_2, B_3, M, E and S. Accordingly, we have the tag sequences S, BE, BB_2E, BB_2B_3E, BB_2B_3ME and $BB_2B_3M\cdots ME$ for characters in a word of length 1, 2, 3, $\cdots$, and 6 (and above), respectively. The six n-gram feature templates are C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1 and $C_{-1}C_1$, where $0, -1$ and 1 stand for the positions of the current, previous and next characters, respectively.

The rule-based segmentation method that we adopt here is the maximum matching algorithm that relies on a predefined lexicon. Given an input sequence, this algorithm tries to find the longest matched word from the lexicon at the current character position. According to the scan direction in the sequence, the algorithm has two variants, forward and backward. We use forward maximum matching (FMM) algorithm in the following experiments.

2.3. Data Splitting

As there are no different sized corpora for any segmentation convention, a series of splitting operations are performed on the original training corpus to generate smaller and smaller training subsets.

To start from the identical size for three selected training corpora, the first two million characters are respectively extracted to build three commensurate training corpora. After that, the two-million-character corpus is split into two equal parts, namely, two training subsets. A CRFs model will be trained from each subset and evaluated in the standard test corpus, and an average F-score will be calculated from both results.[5] And then, each subset will be further split into two parts for CRFs training and test routines. The average F-score is thus calculated from four results. This pipeline of training set splitting, CRFs training and test will be continuously performed until most training subsets include less than one setence. The size of the smallest training subsets is averagely 32 characters, which means that 65,536 training and test routines have to be done for so small subsets.

For a direct comparison, the lexicon for FMM segmentation will be extracted from the same split training subset for CRFs, and the average F-score is also calculated for the same depth splitting.

3. The Results

Figure 1 illustrates two groups of performance curves that represent F-scores for CRFs and FMM methods as the size of training set exponentially varies. Notice that CRFs always outperforms FMM at any size of training corpus.

We observe that the performances given by CRFs linearly increase as training set is doubled. Figures 2(a) shows a linear fitting result for the performance curve on AS corpus. Note that the fitting is quite accurate. This is also the case for the other two corpora, $CityU$ and $MSRA$.

The case of FMM segmentation is sophisticated. We finally found its performance curve can be well represented by the sigmoid function,

$$F = \frac{a}{1 + be^{-x}}, \tag{3}$$

where a and b are two constants, $x = \log(s)$, and s is the size of training data. Figures 2(b) illustrates a good sigmoidal fitting over the performance curves given by FMM segmentation on AS corpus.

The average length of lexicons extracted from training subsets with the same splitting depth is calculated and shown in Figure 3. Again, lines are observed. As either of variant in this figure is in a logarithmic form, this shows that the size of the lexicon is proportional to the power of that of the segmented corpus from where the lexicon extracted. Interestingly, all curves nearly overlap for three different corpora. This suggests that three different segmentation conventions share some similar statistical characteristics In fact, let L denote the size of the lexicon, and s for that of the corpus, we will have

$$L = ks^{0.75} \text{ or } s = k'L^{4/3},$$

for all three corpora according to the results in Figure 3.

Out-of-vocabulary words (OOV) mean those that appear in test corpus but absent in training corpus. They are especially concerned in word segmentation task as the rate of OOV, the proportion of OOV to all words from test corpus, will heavily affect the segmentation performance. And of course OOV is the critical factor which causes data sparseness, thus those unseen items in test corpus could be another reference for our investigation. Figure 4(a) shows how the rate of OOV occurrences or types decreases as the respective training set is continuously doubled (The average strategy is still adopted.). Once more, the OOV rate versus corpus size can be well estimated by a sigmoid function as shown in Figure 4(b) (Figures on the other two corpora are ignored as the similar fitness can be observed, too.).

4. Discussion

On the basis of our empirical results, we can make such a conclusion that the linear growth of performance is at the need of an exponential increasing of corpus size for statistical learning over annotated corpus. This can be roughly explained by Zipf's law, as the learning method is fixed and the data sparsity become worse in an exponential way, the

[4] (Zhao and Kit, 2008b) has shown that this setting is sufficient to output nearly state-of-the-art performances even without a group of additional features derived from unsupervised segmentation. As for CRFs implementation, we use the CRF++ package, http://crfpp.sourceforge.net/

[5] The average operation is especially useful to overcome the uncertainty as the training corpus is extremely small.

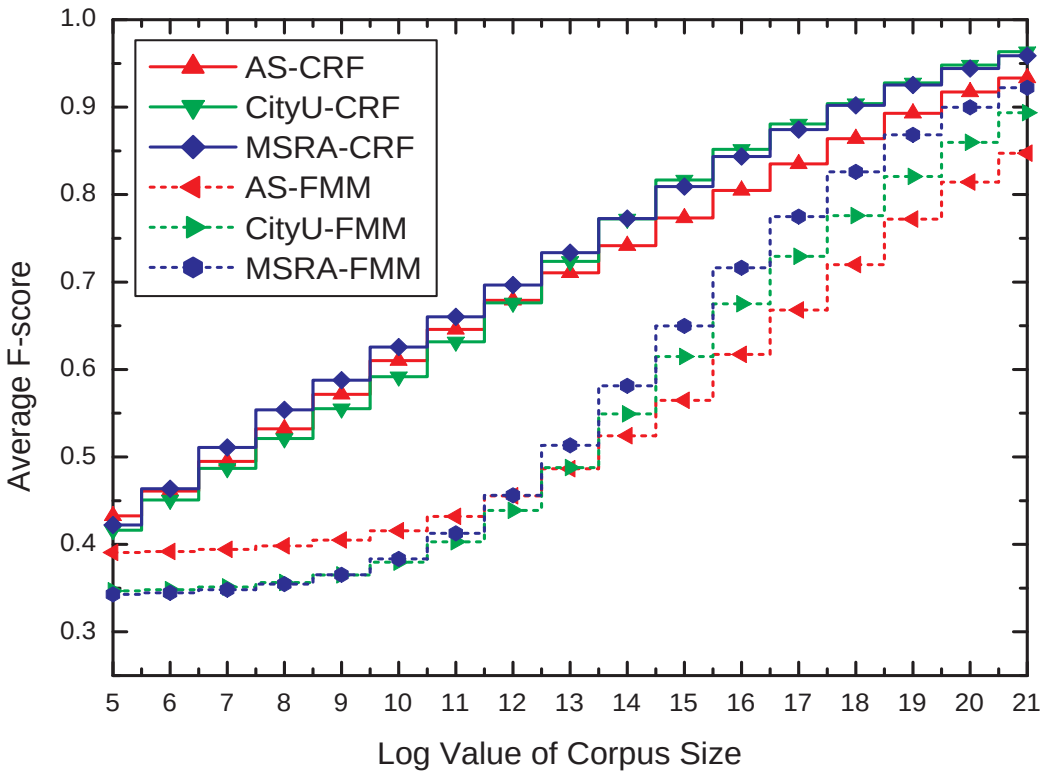


Figure 1: Comparison of learning curves of CRFs and FMM

input data should be exponentially enlarged to overcome such sparsity.

As for a rule-based method, our empirical study shows it is more helpful to overcome data sparsity than a statistical one. Substitute $x = \log(s)$ into equation (3), we have a relation between F-score and corpus size (or lexicon size) as

$$F = \frac{a}{1 - bs} = \frac{a}{1 - b'L^{4/3}} \quad (4)$$

for FMM segmentation, which is actually a negative inverse relation ($a > 0$ and $b, b' > 0$). Though CRFs outperforms FMM in all the above results, this is due to the limited lexicon size for FMM. Note that a lexicon is often much easier to obtain than a segmented corpus, while the former is required by FMM and the latter by CRFs. Thus it will be promising for FMM segmentation as the input lexicon is sufficiently large.

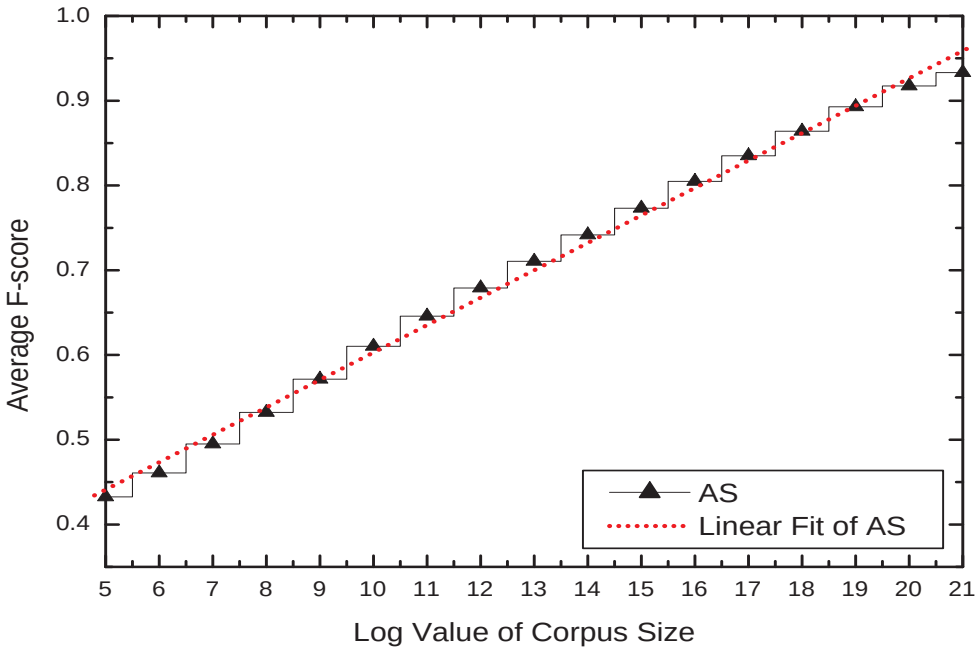
Following the similar way as the above discussion, we say that the OOV rate is in the inverse ratio of the size of training data as in equation (4), the only difference is $b < 0$ at this time. It proves the linear decrease of uncertainty in the data, which helps the linear improvement of the performance, is kept when the corpus is exponentially enlarged.

It is worth noting that this study is actually a computational intensive work, which costs a month to work out nearly 400,000 necessary results of either CRFs or FMM segmentation. Our study on the relations between the learning performance and the data size could provide a good reference for large-scale corpus learning and annotation.

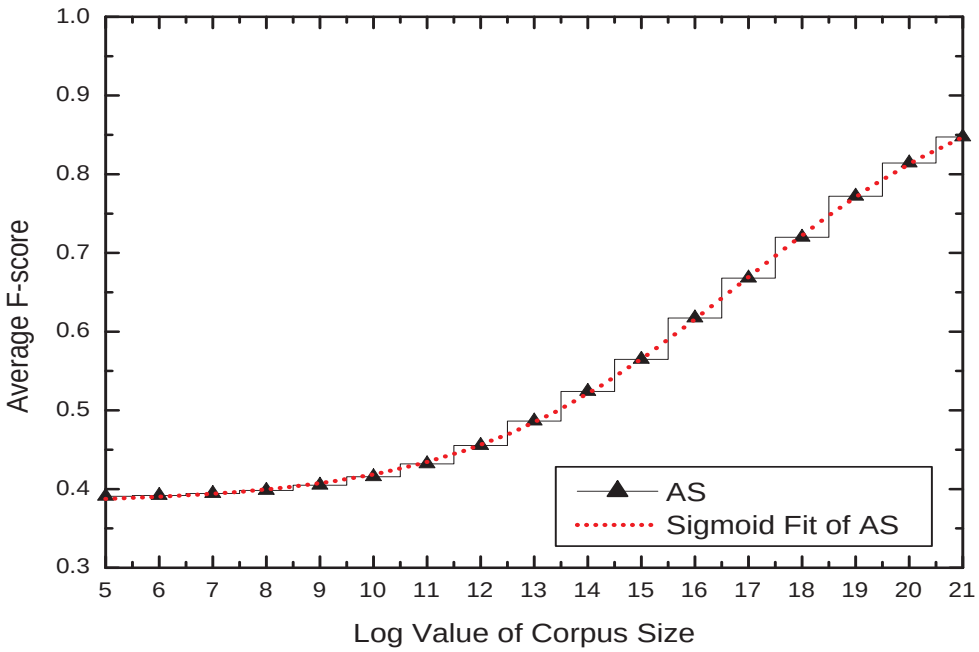
5. Conclusion

Under the empirical investigation with FMM and CRFs based Chinese word segmentation, we report our findings for the performance of different methods with regard of corpus size. As for the choice of our methods, FMM and CRFs are the prevailing representatives of rule-based and statistical approaches for the task of Chinese word segmentation, respectively. Thus our investigation can be described as a generalized case study in computational linguistics, especially for morphological computation.

Of course the overall trend shows that, the larger size of the corpus, the better performance we can obtain, still, when we look into the technical details, we observe that, as for the statistical method, it can be roughly explained by Zipf's law that a linear improvement on performance is actually found with an exponential increasing of corpus size, which can reduce the statistical uncertainty of the low frequent learning examples, as well as OOVs. Meanwhile, interestingly, the rule-based method shows its strength in terms of data sparseness, that can be ascribed to those learning examples with regardless of statistical significance are still well learnt in this kind of learning strategy.



(a)



(b)

Figure 2: Performance curves with fitting: AS corpus ((a) CRFs, (b) FMM)

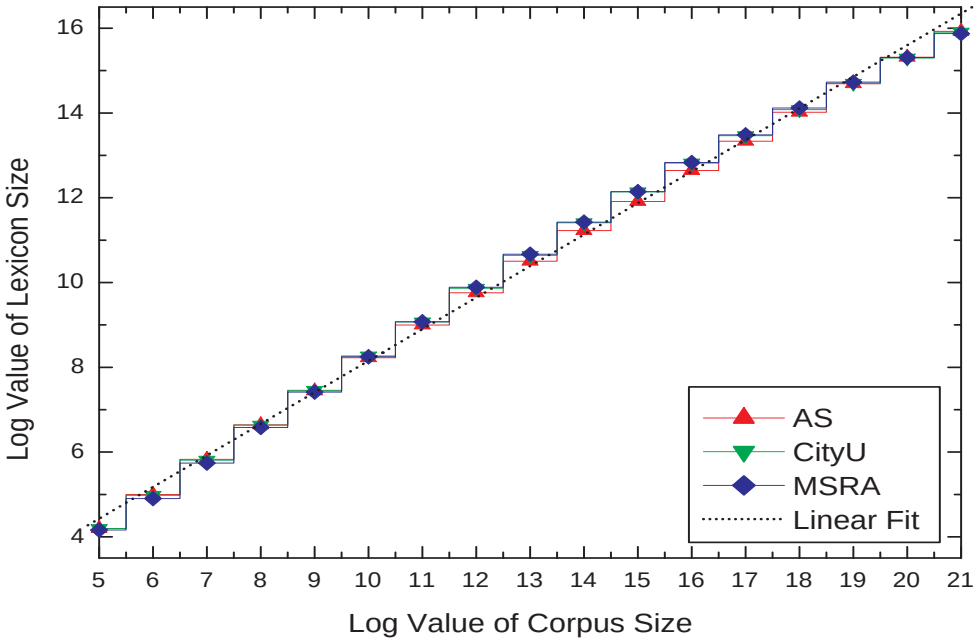


Figure 3: lexicon size vs. corpus size

Moreover, the relationships that we have demonstrated in this paper do give us some hints about constructing a balanced corpus, and on the other hand, we may find a promising way of joint approach with rule-based and statistical methods incorporated for a certain task on a certain sized corpus.

Acknowledgments

The research described in this paper was partially supported by the National Natural Science Foundation of China (Grant No. 60903119), and partially supported by City University of Hong Kong through the Strategic Research Grant 7002267 and 7002388.

6. References

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *ACL-2001*, pages 26–33, Toulouse, France.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia, July 22-23.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *COLING 2004*, pages 562–568, Geneva, Switzerland, August 23-27.

Yan Song, Jiaqing Guo, and Dongfeng Cai. 2006. Chinese word segmentation based on an approach of maximum entropy modeling. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 201–204, Sydney, Australia, July.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

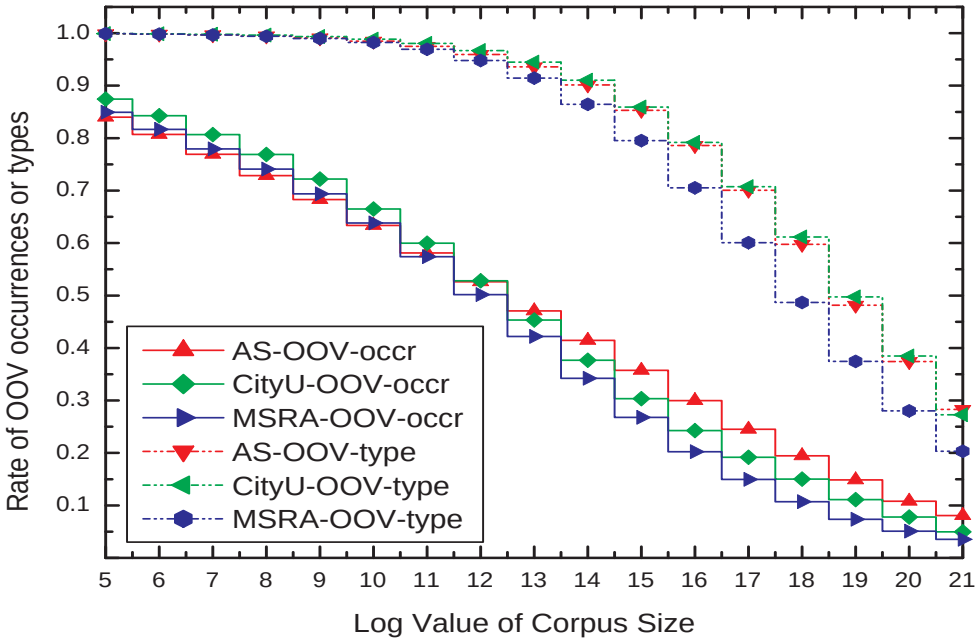
Hai Zhao and Chunyu Kit. 2008a. Exploiting unlabeled text with different unsupervised segmentation criteria for chinese word segmentation. In *Research in Computing Science*, volume 33, pages 93–104.

Hai Zhao and Chunyu Kit. 2008b. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *The Sixth SIGHAN Workshop on Chinese Language Processing*, pages 106–111, Hyderabad, India, January 11-12.

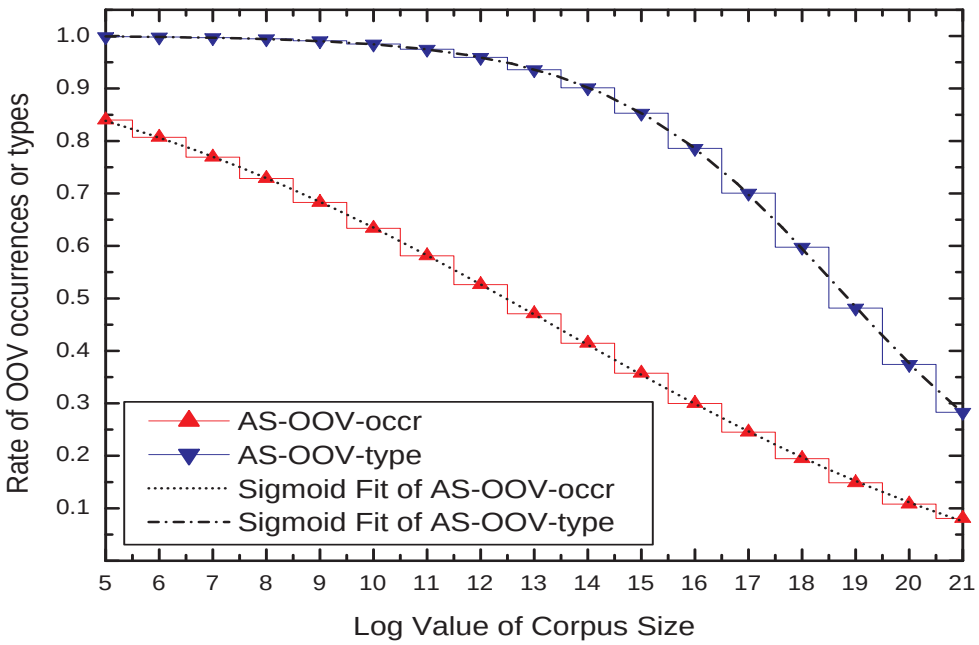
Hai Zhao, Chang-Ning Huang, and Mu Li. 2006a. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165, Sydney, Australia, July 22-23.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006b. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In *Proceedings of the 20th Asian Pacific Conference on Language, Information and Computation*, pages 87–94, Wuhan, China, November 1-3.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.



(a)



(b)

Figure 4: (a) OOV rate vs. corpus size, (b) OOV rates with sigmoidal fitting: AS corpus