Why Chinese Web-as-Corpus is *Wacky*?
Or: How Big Data is Killing Chinese Corpus Linguistics

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

This paper aims to examine and evaluate the current development of using Web-as-Corpus (WaC) paradigm in Chinese corpus linguistics. I will argue that the unstable notion of wordhood in Chinese and the resulting diverse ideas of implementing word segmentation systems have posed great challenges for those who are keen on building web-scaled corpus data. Two lexical measures are proposed to illustrate the issues and methodological discussions are provided.

Keywords: Corpus evaluation, word segmentation, Web as Corpus

1. Introduction

The emergence of big data has brought about a new paradigm shift to all the fields related to data analysis. As one of the tenets of corpus linguistics is to collect authentic texts for empirical linguistic analysis, it is not surprising that, with the explosion of the massive and diverse web data and processing tools increasingly available like never before[1], corpus linguistics over the recent years has witnessed a dramatic change of research paradigm, too. In a positive way, the easy availability and unique potential to yield large-volume linguistic data on up-to-date language use from the web democratize the way linguists work, and liberate the creativity in studying the intricacy of language. Enthusiastic attitude toward using web as a source of corpus data has been around for many years, and the term *Web-as-Corpus* (WaC) has been taken to refer to various approaches of exploiting the Web for linguistic studies (Kilgarriff and Grefenstette, 2003).

Constructing a large corpus from the web either from scratch, or as one additional resource to complement the existing compiled corpus, has also become one of the most prosperous on-going work in the community of Chinese corpus linguistics. It is, however, challenging and difficult not only because of the general issues such as *ephemeral nature* of the web, *replicability* or *reliability* of the results, but also due to a more methodological problem relating the interplay with *corpus size* and *word segmentation*. General procedure in constructing a web corpus involves the crawling, pre-processing and annotation of the data. Among the pre-processing tasks in particular in Chinese, **word segmentation** as a specific kind of tokenization is normally required to perform on the cleaned raw data. Even though the current segmentation systems achieve great performance up to over than 90% percent accuracy given a standardized segmentation scheme, the errors increase proportionally with the increase of corpus size. It get even worse as scaled corpus data keep increasing in size, there will be even less chance to verify the reliability via human intervention. This leaves the problem unsolved. This study thus aims to pose the thorny issues, and in light of this, propose a series of evaluative measures in the hope to stimulate further reflections on the role of Big Data in the context of building web corpus for languages whose wordhood can only be *functionally* defined.

2. Review of Web Corpus Development

Although the WaC has redefined many ways of research methods in linguistics, there is yet no unified understanding of how it is defined. For example, the web can be used as the data source, or it can be turned into an interface of linguistically-oriented meta-search engine like `WebCorp`[2]. In spite of its popularity, many have questioned WaC regarding its *representatives* and *reliability* (Kilgarriff and Grefenstette, 2003). The **evaluation** of the WaC, in particular, in comparison with common corpora has become a crucial task. After experimenting with different proposed measures of corpus similarity, (Kilgarriff, 2001) found that χ^2 outperforms others and is shown to be the best measure both in measuring the similarity of a corpus to itself and cross-corpora comparison. However, this measure is not *text-length invariant* and thus not suitable for comparing with scaled data. In addition, such word-based measure is particularly hampered by the lack of *stable* notion of word as counting unit in the case of Chinese WaC.

As well-known in the field of Chinese NLP, one of the most challenging tasks in preprocessing corpus in Chinese is the word segmentation, for there is no natural indicator for the word boundary in the running texts. Among different segmentation algorithms, the lexicon-based approach is widely adopted, which can to a great extent identify Chinese sentences as distinct words from Chinese texts. However, the word identification ability of the lexicon-based scheme is highly dependent on a well-prepared lexicon with sufficient amount of lexical entries. Hybrid approach thus proposed to combine with other statistical information to detect out-of-vocabulary (OOV). Notwithstanding the tremendous efforts to handle with this issue, it is not possible yet to achieve a commonly accepted solution, either linguistically or technically. Based on the background, this paper thus

[1] e.g., `BootCat toolkit` at http://bootcat.sslmit.unibo.it

[2] http://www.webcorp.org.uk

tends to be more cautious, focussing as much on the potential as on the hazards of using the web as corpus in Chinese context. The following section will use the WaC we built as an example to illustrate the crucial issues, and call for a sound evaluative methodology.

3. Building and Evaluating Chinese Web Corpus

Since 2009 we have been building a Taiwanese Mandarin WaC, with the goal of complementing the Academia Sinica Balanced Corpus of Modern Chinese (henceforth ASBC) with a corpus from the web. Microblogs (Plurk (`http://www.plurk.com`)) was chosen because it reflects the current uses of language on the one hand, and the analysis of *microtext* has potential applications in many aspects. Currently, our corpus contains 3000 plurkers, with *plurk data objects* (e.g., plurk_id, date of the plurk posted, content, response, emoticons, user's meta-information. etc) which amounts to 15 GB in size.

After automatically collecting and preprocessing the massive plurk data, the problem encountered soon after that stage was the lack of established and methodologically-sound measures for homogeneity for Chinese corpora. As already mentioned in 2. and 3., the proposed measure such as χ^2, `Mann-Whitney` ranks test and other frequency profiling that work well in English would not be suitable for Chinese WaC. It is rather knotty to make the comparison feasible.[3] To handle with text-length dependence as well as the effect of segmentation errors, two measures of lexical statistics from distributional perspective are proposed as our starting point.

3.1. Lexical Richness

Frequency distribution analysis has played an important role in corpus linguistics. In particular, relative frequency counts - e.g., frequencies per 10 million words - is of fundamental importance when we want to compare linguistic patterns from different corpora or different subsections of a corpus. (Baayen, 2001) proposes the notion of **frequency spectrum** that provides a concise summary of a frequency distribution of corpus data.[4] Frequency spectrum uses the symbol V for vocabulary size (number of types), N for sample size (number of tokens) and V_m (m an integer value) for the number of types that have frequency m. In particular, V_1 is the number of *hapax legomena*, i.e., the number of types that occur only once in the corpus. Figure 1 shows a contrastive plot of the first 50 spectrum elements with the x (i.e., m) axis on a logarithmic scale of four corpora: ASBC (Academia Sinica Balanced Corpus), Plurk corpus, Brown corpus and LDC Chinese Gigaword Corpus.[5]

Frequency spectrum as shown is characterized by "very high values corresponding to the lowest frequency classes, and a very long tail of frequency classes with only one

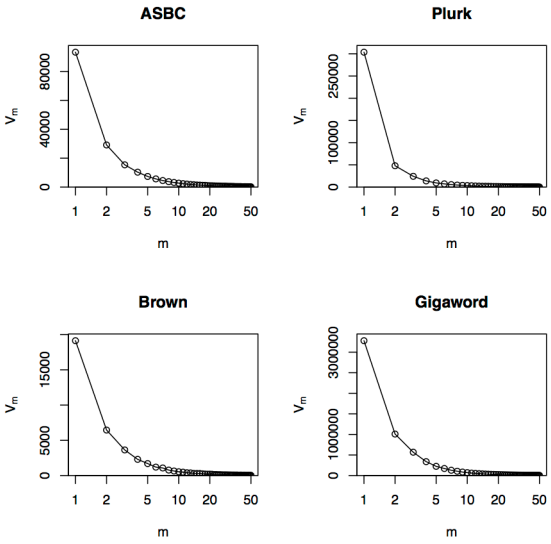

Figure 1: Frequency Spectra of Different Corpora

member" among the four corpora. It actually provides *probabilistic* information of encountering new types if we were to sample more data of the same category. In order to get deeper understanding of how rapidly vocabulary size is growing with increasing size, we use another data structure called **vocabulary growth curve** (`vgc`), which reports vocabulary size (number of types, $\mathcal{V}$) as a function of sample size (number of tokens, $\mathcal{N}$). It is estimated by the "ratio of the number of hapax" (types with a frequency of 1) to the number of tokens sampled. The growth rate is a probability, the probability that, after having read $\mathcal{N}$ tokens, the next token sampled represents an unseen type, a word type that did not occur among the preceding $\mathcal{N}$ tokens (Baayen, 2001). Take ASBC for instance, the first few rows of `vgc` object are:

```
     N    V  V1
1 1000  552 392
2 2000  867 561
3 3000 1095 677
```

This means that, after the first 1,000 tokens in the ASBC, we saw 552 distinct types, and 392 of them being *hapax legomena* (i.e., having occurred only once) at that point, etc. A comparative vocabulary growth plot with $\mathcal{V}$ and $\mathcal{V}_1$ curves is shown in Figure 2. We can discern the class differences in Figure 2. For ASBC and Brown corpus where segmentation is not an issue or resolved, the curves smooth and relatively lessen at some points; while for Plurk and Gigaword gigantic corpus, $\mathcal{V}_1$ seems to keep increasing beyond expectation.

3.2. Lexical Coverage

Another distributional measure we propose to compare Chinese WaC with traditional corpus is called *lexical coverage*. *Lexical coverage* generally refers to the percentage of running words in the text that readers understand (Nation, 2006). In the field of language learning and readabil-

[3] (Tang and Chen, 2011) compiled a plurk corpus and compares it with ASBC in terms of word frequency, lexical semantics and sentiment expression.

[4] The freely available and easy-to-use `zipfR` package provides various functions to explore the spectrum objects. `http://zipfr.r-forge.r-project.org/`.

[5] http://catalog.ldc.upenn.edu/LDC2007T03

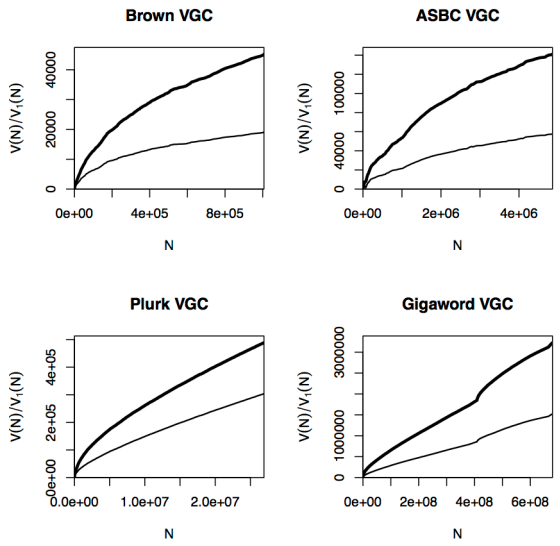

Figure 2: Vocabulary Growth Cruves of Different Corpora

ity studies, the rationale behind the lexical coverage is to measure the degree of unknown vocabulary which can be tolerated in a text before it interferes with comprehension. In many Natural Language Processing (NLP) applications where lexical resources are heavily relied upon, the estimate of the proportion of *Out-of-Vocabulary* (OOV) is a crucial work as well. In the following, we estimate the lexical coverage to compare different corpora with the aims to see how "segmentation errors" affect Chinese WaC.[6]

For the sake of cross language comparison, the first 100k lemma tokens are extracted from the Brown corpus, ASBC and Plurk corpus, respectively. For the first 100k lemma, the vocabulary size for each of the three corpora is (12780, 15613, 20050). By subtracting the *hapax legomenon*, we get (6477, 7463, 7057). In this way, we see that the percentages that count as OOV types in the three samples are around (50%, 52%, 63%), and the proportions of the overall tokens they account for are (6%, 8%, 12%). That is, we can clearly see that the OOV in Plurk corpus outnumbers the other two traditional corpora both in type and token proportion.

Given the vocabulary size and the frequency spectrum at this sample size, (Baayen, 2001) proposes that we can work backwards to smaller sample sizes (interpolation) and forwards to larger sample sizes (extrapolation) to get smoother curve. In order to observe whether the distribution $\mathcal{V}_1$ or $\mathcal{V}$ is an indicator for the evaluation of Chinese WaC, we can *extrapolate* $\mathcal{V}$ to larger samples by resorting to a family of statistical models for the word frequency distribution: **LNRE** (Large-Number-of-Rare-Events) models. It is argued in (Baayen, 2001) that word frequency distributions are LNRE distributions, which is characterized by the presence of large numbers of words with very low probabilities of occurrences. Among different models of LNRE introduced in (Baayen, 2001), three models clear outperform the others (Evert and Baroni, 2005): *Generalize Inverse Gauss Poisson*, *Zipf Mandelbrot* and *finite Zipf Mandelbrot*.

[6]Comparison of lexical coverage in Taiwan Mandarin conversation and a balanced corpus can be referred to (Tseng, 2013)

		N1	N10	N100
Brown	exp.$\mathcal{V}$	24951.19	25567.39	25567.39
	exp.OOVs	0.7404132	0.7466695	0.7466695
ASBC	exp.$\mathcal{V}$	34988.97	37246.81	37246.81
	exp.OOVs	0.7867042	0.7996339	0.7996339
Plurk	exp.$\mathcal{V}$	69223.01	99156.87	99251.33
	exp.OOVs	0.8980541	0.9288299	0.9288977

Table 1: Comparison of Brown, Plurk and ASBC based on LNRE model

In the case of corpora comparison, we use a LNRE (finite Zipf-Mandelbrot) model from the available frequency spectra of three corpora to estimate the expected proportion of OOV types and tokens when larger $\mathcal{N}$s is extrapolated. The following shows the details of LNRE modeling on ASBC 100k corpus.

```
finite Zipf-Mandelbrot LNRE model.
Parameters:
   Shape:          alpha = 0.782017
   Lower cutoff:       A = 1.503985e-06
   Upper cutoff:       B = 0.005448517
 [ Normalization:      C = 0.8156926 ]
Population size: S = 37246.81
Sampling method: Poisson, with exact calculations.

Parameters estimated from sample of size N = 1e+05:
                 V      V1      V2      V3     V4    V5
   Observed: 15613 8150.00 2463.00 1159.00 753.00 504.0 ...
   Expected: 15613 8183.75 2779.59 1223.13 681.78 438.9 ...
   Goodness-of-fit (multivariate chi-squared test):
          X2 df           p
     399.726 13 2.555398e-77
```

Based on this estimated model, we can get the proportion of expected $\mathcal{V}$s at arbitrary $\mathcal{N}$s. Assuming that we have $\mathcal{N}$s of 1,10,100 million tokens, the expected values of $\mathcal{V}$ and the proportions of OOV types for each corpora are listed in Table 1. It is clearly shown that the increasing corpus size (N1, N10, N100) would result in the *stable* equilibrium point of the proportion of OOVs in traditional corpora (Brown and ASBC), while not in the case in Plurk corpus.

4. Discussion

As is observed in (Baayen, 2001), "word frequency distributions generally have a large growth rate even at the full sample size, implying that there are more types to be sampled if more word tokens are added to the sample". But it is also observed, as can be seen in Figure 2, the vocabulary growth curve (in terms of OOV rate) will necessarily be increasing, and will normally be negatively accelerated (that is, its rate of increase will slow down) in Brown corpus, where word boundary is already delimited, and ASBC, where word boundary delimitation that is automatically by segmentation system is evaluated and corrected by humans.

	Plurk corpus	ASBC	ratio
hapax legomena (V1)	303,629	93,306	3.25
sample size ($\mathcal{N}$)	26,930,077	9,252,220	2.91
vocabulary size ($\mathcal{V}$)	488,531	219,659	2.22
V2	48,035	29,128	1.64
V3	24,275	15,402	1.58
V4	13,950	10,314	**1.35**
V5	9,371	7,349	**1.28**

Table 2: Ration Comparison of Plurk and ASBC

While in scaled WaC like Plurk corpus, in which comprehensive human evaluation is impossible, the curve behaves beyond the expectation. It also holds true for the lexical coverage.

Another interesting issue can be seen in Table 2. The high proportion of hapax (V_1) in Plurk corpus arguably explains the errors caused by segmentation system. The ratio of difference of V_2 to V_5 decreases and gets closer to one. This might lead one to assume that we can use V_5 for instance, as the threshold in building the lexicon, and removing the sentences where V_1 to V_4 occurs. But a closer inspection of corpus data shows that it is still problematic due to the *non-representativeness* of web genre, which results in the highly repeated similar errors.

5. Conclusion

Over the recent years, the potentials of Web as corpus have been widely recognized in corpus community, for it offers a multitude of possibilities for corpus research, and with the rapid development of cloud-based infrastructure, we believe that using web as corpus and harvesting the web for linguistic purposes will soon delve into an important subfield of corpus and computational linguistics. However, under the lens of lexical statistic analysis, as cautioned in this paper, the focus should be turned to the evaluative methodology when facing with BIG data in the context of Chinese WaC, due to the effect of word segmentation. A resort to collective intelligence with reproducible architecture is envisioned.

6. Acknowledgements

The author would like to acknowledge the assistance of Mei-Yu Chen for this work. The author would also like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of this paper.

7. References

Baayen, R. H. (2001). *Word frequency distributions*, volume 18. MIT Press.

Evert, S. and Baroni, M. (2005). Testing the extrapolation quality of word frequency models. In *Proceedings of Corpus Linguistics*, volume 2006.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133.

Nation, I. S. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La revue canadienne des langues vivantes*, 63(1):59–82.

Tang, YiJie, C.-Y. L. and Chen, H.-H. (2011). A comparison between microblog corpus and balanced corpus a comparison between microblog corpus and balanced corpus from linguistic and sentimental perspectives. In *Analyzing Microtext*.

Tseng, S.-C. (2013). Lexical coverage in taiwan mandarin conversation. *Computational Linguistics and Chinese Language Processing*, 18(1):1–18.