N-gram Statistics in English and Chinese: Similarities and Differences

Stewart Yang, Hongjun Zhu, Ariel Apostoli and Pei Cao
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043

Abstract

Chinese and English belong to two very different families of human languages. Yet, since the underlying human concepts are universal, one can expect that there are many statistical similarities between Chinese texts and English texts. In this paper, we present results of analyzing the quantity and frequency of N-grams in 200 million randomly-sampled English and Chinese web pages.

The similarities and differences in N-gram frequency distributions yield important insights about the two languages. First, the distribution of the unique number of N-grams is similar between English and Chinese, yet the Chinese distribution is "shifted" to larger N. The distribution indicates that on average, 1.5 Chinese characters correspond to 1 English word. Second, while frequency distributions of uni-grams and bi-grams are very different between Chinese and English, the frequency distribution for 3-grams and 4-grams are strikingly similar between Chinese and English. This leads to the conjecture that in both languages, frequent 3-grams and 4-grams represent the same set of concepts and patterns.

1. Introduction

Given a large enough collection of documents, what can one learn about the frequently occurring sequences of N words (also called N-grams)? Do they merely represent pattern of speech or do they consist of commonly occurring concepts and entities? For languages as different as English and Chinese, do characteristics of the N-grams differ significantly?

In this paper, we attempt to obtain some preliminary answers to the above questions by processing a set of 200 million web pages. The Web gives one easy access to many documents authored by different individuals. We chose the web pages randomly from the repository of web pages crawled by Google. For each sentence in each web page, all sequences of N words (*N-grams*) are extracted, with N varying from 1 to 30. We then study the characteristics of frequently appearing N-grams.

This study covers two languages: English and Chinese. The two belong to very different families of languages; hence it's interesting to compare statistics about them. For this study, 100 million English web pages and 100 million Chinese web pages are used.

A major difference between English texts and Chinese texts is the lack of spaces in Chinese texts. English words are separated by spaces in a sentence, but Chinese words are not. Chinese words that correspond to English words may contain multiple Chinese characters. Segmentation, i.e. partitioning a Chinese sentence into a sequence of words that correspond to English words, tends to be a challenging task.

We ignore the issue of segmentation completely when processing Chinese documents. Instead, each Chinese character is treated as a "word" when constructing sequences of N words. As a result, a uni-gram (i.e. 1-gram) in Chinese does not necessarily correspond to a uni-gram in English. However, the distinction blurs as N is increased.

The results of the study are as follows:

- The trend of the total number of unique N-grams as a function of N is similar in English and Chinese, but the Chinese version is shifted to the right. The curves indicate that, on average, 1.5 Chinese characters correspond to 1 English word.

- While the total number of unique N-grams is higher in English than in Chinese for $N < 5$, this is no longer the case when we limit the N-grams to those that appear at least 5 or 10 times in the corpora. This indicates that the impact of typos is higher in English than in Chinese.

- The frequency distributions of the 100,000 most popular uni-grams in Chinese shows a distinct knee around 500, which is missing in the distribution for English. This indicates that the most common 500 Chinese characters are used far more frequently in composing words than the others.

0-7695-2997-6/07 $25.00 © 2007 IEEE
DOI 10.1109/ICSC.2007.46

- English and Chinese have nearly identical frequency distributions of the 100,000 most popular 3-grams and 4-grams. One possible explanation is that 3-grams and 4-grams in both languages represent the same set of concepts and entities.

In summary, though simplistic, N-gram statistics can yield surprising insights into languages when used on a large enough corpora.

2. Processing of Web Pages

We leveraged the computing infrastructure at Google to process the large amount of documents. Two MapReduce [1] steps are used to generate the statistics.

In the first MapReduce step, all text "chunks" from all web pages are extracted and aggregated. The Map phase performs the extraction. For each web page, the text is broken into "chunks" by breaking at major punctuation symbol and major HTML tag. The chunks are output into the Reduce phase. The reduce phase gathers all the text chunks, sorts them, and counts the appearance of each text chunk. The output file, which is a collection of $< textchunk, count >$ pairs, is fed into the second MapReduce step.

In the second MapReduce step, all N-grams are extracted from the text chunks and aggregated. In the map phase, for each text chunk of length l, all sequences of N words, together with the count of the text chunk, are output. The Reduce phase sums all the counts for each N-gram and outputs a list of $< N - gram, count >$ pairs. The list is then sorted and aggregated to generate statistics.

3. Number of Unique N-Grams

The first set of statistics is on the number of unique N-grams. Since many of the unique N-grams are due to typos, we also examine the number of unique N-grams for N-grams that appear at least 5 or 10 times in the corpora.

Figure 1 shows the total number of unique N-grams for N from 1 to 30. In both languages, as N increases the number of unique N-grams goes up initially and then goes down. The reduction is mostly due to limits on the length of text chunks, since most text chunks are short.

As N increases, the increase in the number of unique N-grams decelerates. The increase is most drastic from 1-grams to 2-grams: a factor of 6.5 for English, and a factor of 5.2 for Chinese. The increase from 2-grams to 3-grams is also significant: a factor of 4 for English and a factor of 5 for Chinese. However, as N increases further the increase in the number of unique N-grams is less than a factor of 2.

In other words, in a large collection of web pages, the probability that an arbitrary uni-gram A and an arbitrary uni-gram B appears together is very low, on average. Furthermore, for an arbitrary 3-gram C, on average, there are less than 2 unique succeeding words in the corpora.

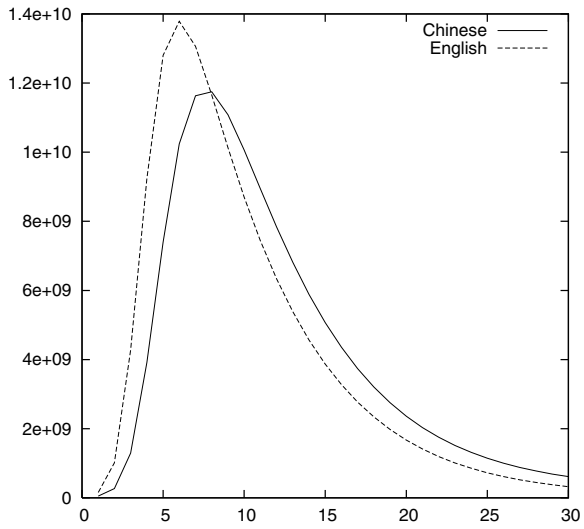

Figure 1. Total number of unique N-grams, for all N-grams.

Figure 2 and Figure 3 are statistics when we only consider N-grams that appear at least 5 times and 10 times, respectively. They shows similar trends as N increases.

Comparing the three figures, one can see that when all N-grams are counted, English has more N-grams than Chinese for small N. However, when only N-grams that appear at least 5 or 10 times in the corpora are counted, English has fewer N-grams. One possible explanation is that the N-grams that appear less than 5 times in the corpora are typos or very unique names, and it's conceivable that English has more typos and unique names than Chinese.

The number of N-grams peak at different N for English and Chinese. If all N-grams are considered, the total number of unique N-grams peak at 6-grams in English and around 9-grams in Chinese. If only N-grams that appear at least 10 times in the corpora are considered, the peak is 4-grams in English and around 6-grams in Chinese. In other words, a rough rule of thumb is that 1.5 Chinese characters corresponds to 1 English word.

4. Frequency Distribution of Popular N-Grams

Do the frequency distributions of popular N-grams in English and Chinese follow a power law distribution, often called Zipf's law [5]? Are there differences in the distribution between English and Chinese?

To answer these questions, we plot the frequency distri-

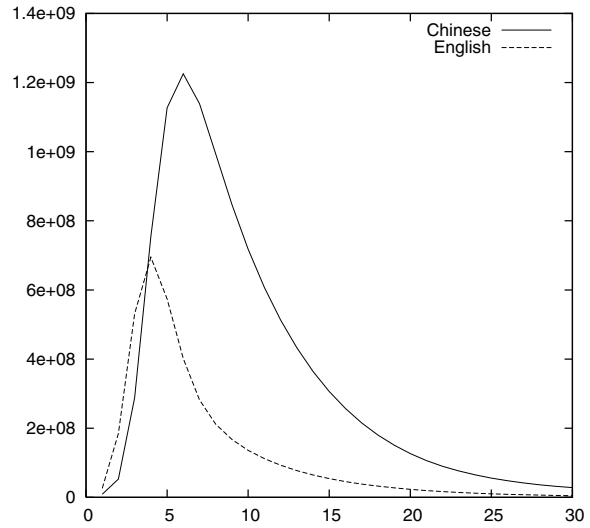

Figure 2. Total number of unique N-grams, for N-grams that appear at least 5 times in the corpora.

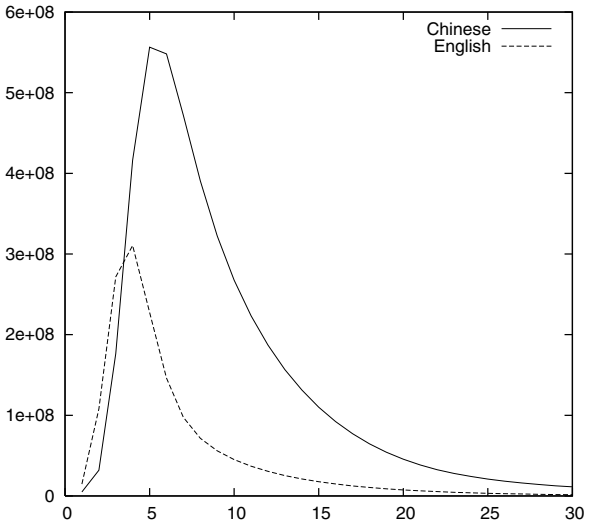

Figure 3. Total number of unique N-grams, for N-grams that appear at least 10 times in the corpora.

bution of the most popular 100,000 N-grams, for N from 1 to 6. For each N, we order the N-grams from the most popular to the least popular, and plot the frequency of the k'th most popular N-gram as a function of k. We show two figures, one in which the x-axis is in linear scale and one in which the x-axis is in log scale. The former shows us the trend for the "body" of the popular N-grams, i.e, from 10,000-th N-gram to 100,000-th N-gram. The latter shows us the trend for the "head", i.e. the most popular 10,000 N-grams. In both figures, the y-axis is in log scale, that is, for a fixed N the y-axis is $log_2(frequency of N-gram / total frequency of all N-grams)$.

Figure 4 show the results on uni-grams. The distribution is "shallower" in English than in Chinese. In other words, in Chinese a fewer number of uni-grams/characters account for more appearances than in English. This is understandable because a Chinese character is often not a word, but only part of a word. In fact, there is a clear knee in the distribution for Chinese uni-grams: around 500-1000. Thus, the top 1000 commonly used Chinese characters accounted for a large percentage of Chinese characters in the web pages. We speculate that these characters are easy to write and easy to input.

Figure 5 shows the results on bi-grams. The distributions of the two languages become closer. Both language show a knee around 100.

Figure 6 shows the results on 3-grams. Except for the top 100 3-grams, the distribution is nearly identical for Chinese and English. This result is a surprise! Figure 7 shows the results on 4-grams. Again, except for the top 100 entries, the distribution is very close in English and in Chinese.

The results suggest that, while 1-grams and 2-grams are fundamentally different in Chinese and English, 3-grams and 4-grams are of similar enough nature that the distributions are the same in both languages. Thus, we speculate that the 3-grams and 4-grams tend to represent independent units of expressions or concepts in both languages.

Figure 8 and Figure 9 present results on 5-grams and 6-grams. There are minor differences between English and Chinese in the distribution. The Chinese distributions are slightly flatter than the English ones, though the trends are still quite close.

5. Analysis of Popular N-Grams

Do the popular N-Grams represent valid concepts or merely common patterns of speech? To answer this question we manually examine the top 100,000 multi-grams (3-grams, 4-grams and 5-grams) in both English and Chinese.

The top 100 3-grams in both languages are mostly phrases specific to the Web environment. For example, the top 3-gram in English is "all rights reserved", not surprising since most corporate web pages have the phrase at the

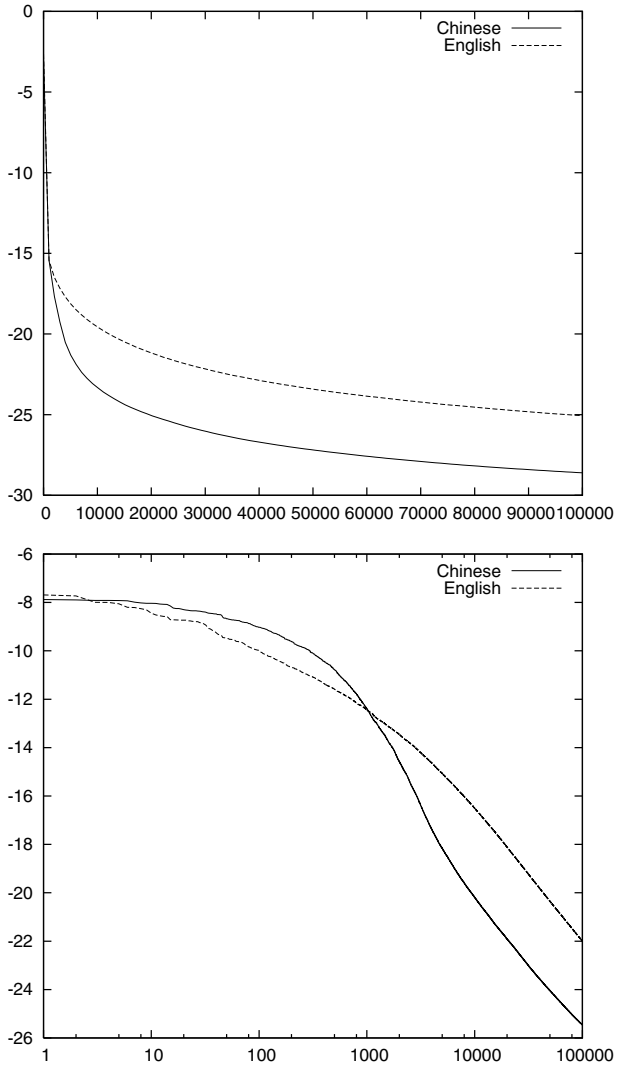

Figure 4. Frequency distribution of the 100,000 most popular uni-grams, with the x-axis in both linear scale and log scale. The y-axis is in log scale in both figures.

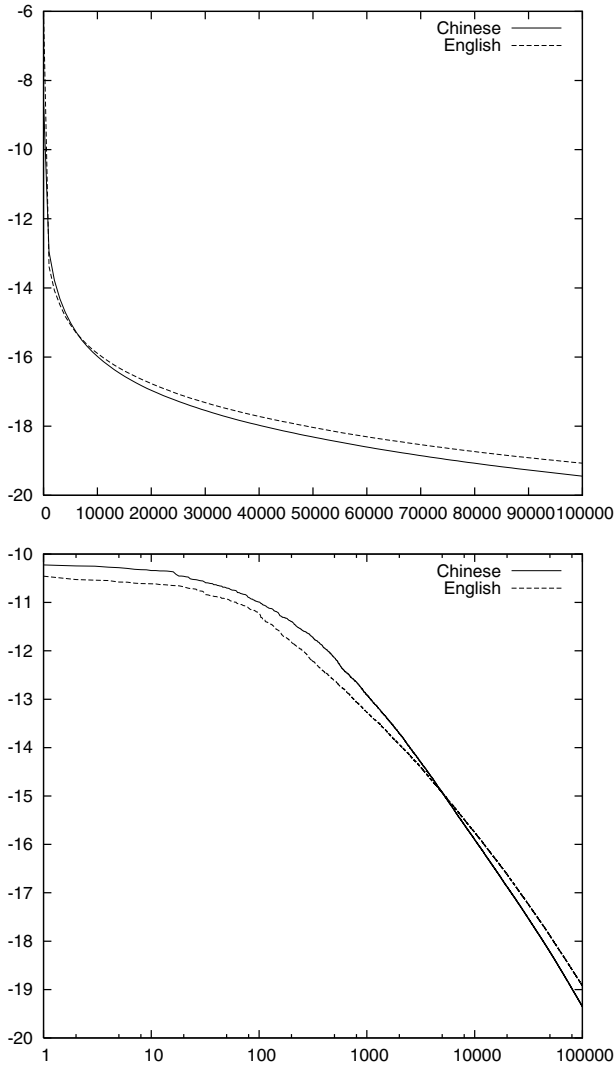

Figure 5. Frequency distribution of the 100,000 most popular bi-grams, with the x-axis in both linear scale and log scale. The y-axis is in log scale in both figures.

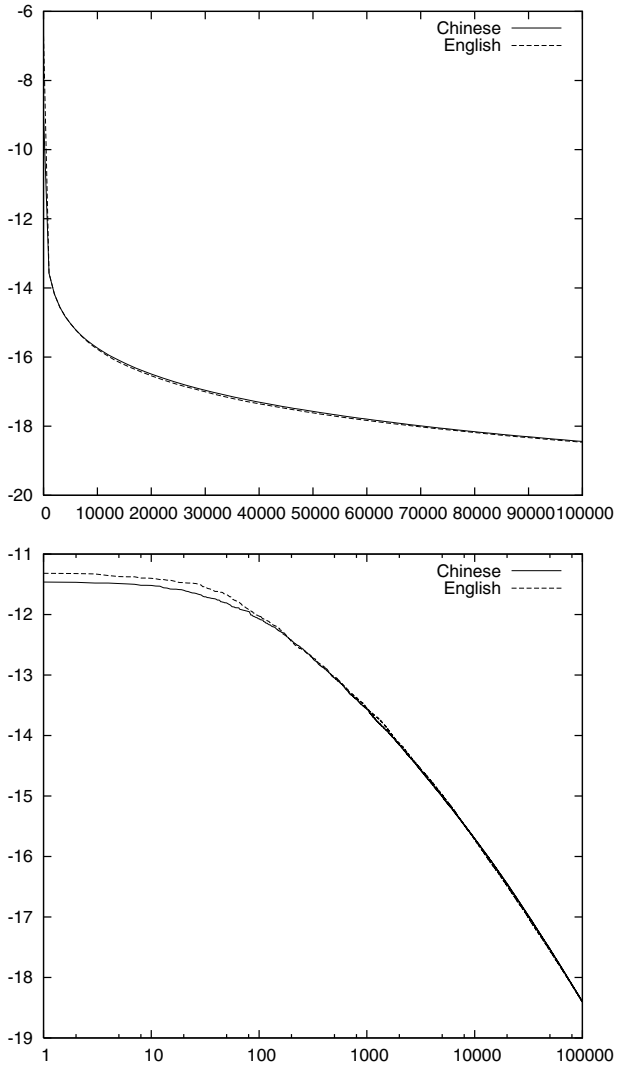

Figure 6. Frequency distribution of the 100,000 most popular 3-grams, with the x-axis in both linear scale and log scale. The y-axis is in log scale in both figures.

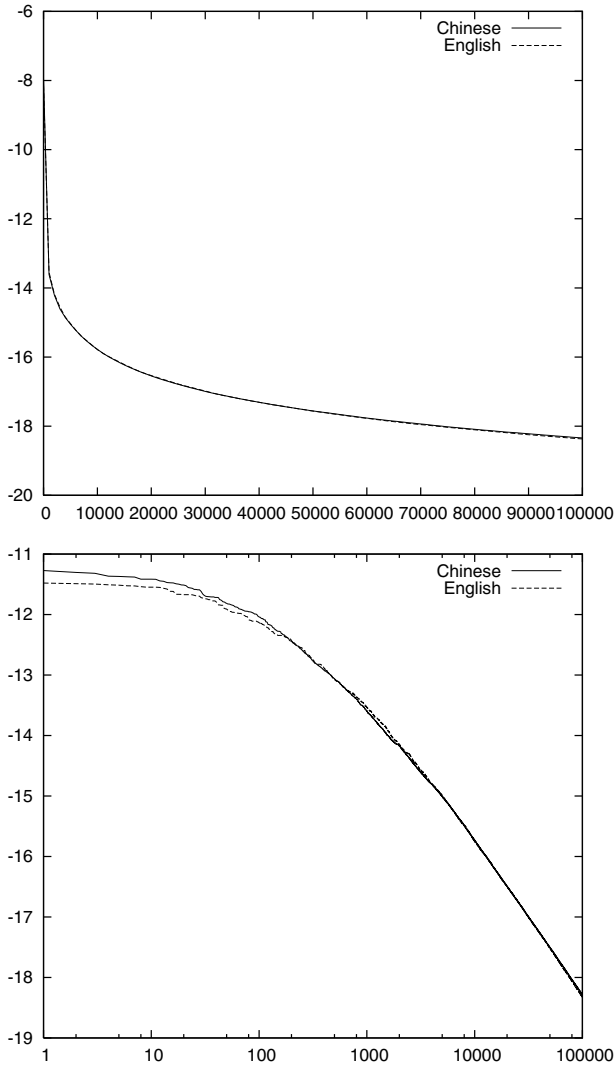

Figure 7. Frequency distribution of the 100,000 most popular 4-grams, with the x-axis in both linear scale and log scale. The y-axis is in log scale in both figures.

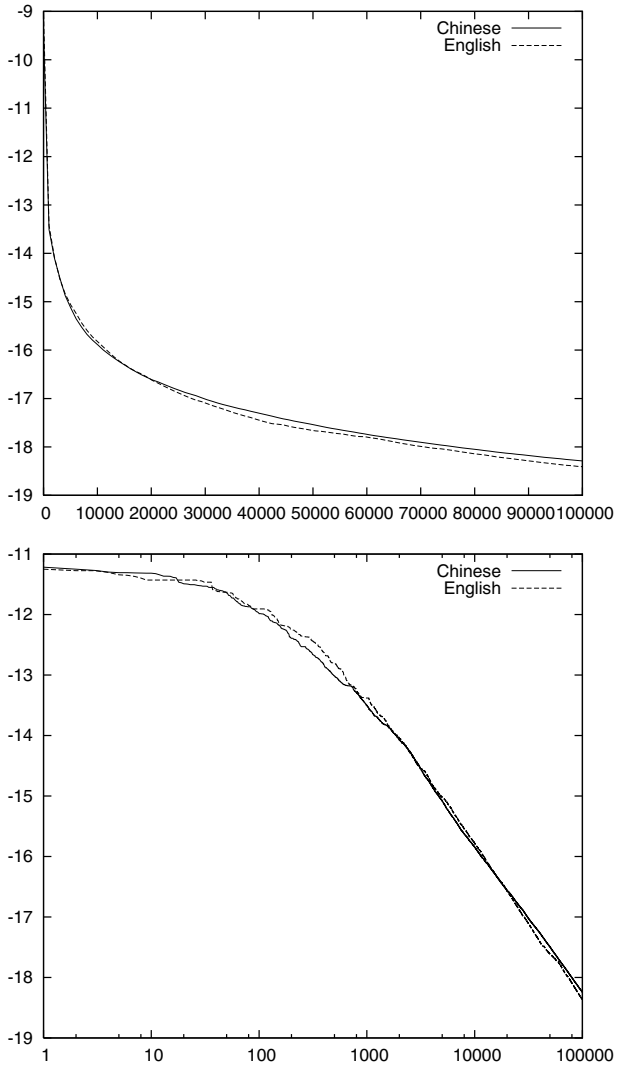

Figure 8. Frequency distribution of the 100,000 most popular 5-grams, with the x-axis in both linear scale and log scale. The y-axis is in log scale in both figures.

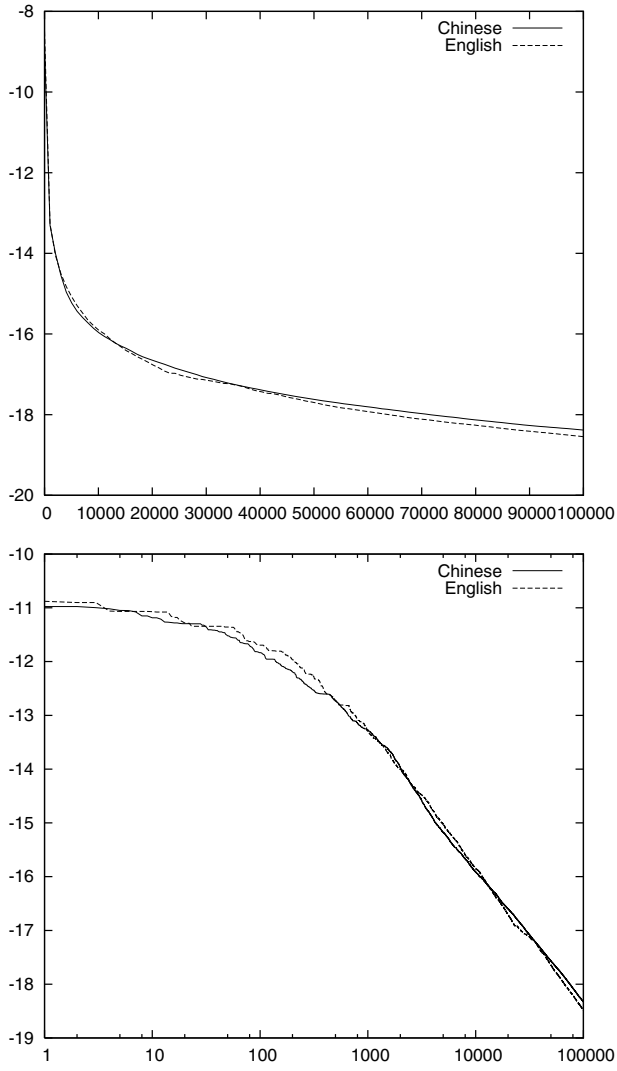

Figure 9. Frequency distribution of the 100,000 most popular 6-grams, with the x-axis in both linear scale and log scale. The y-axis is in log scale in both figures.

bottom. The top multi-gram in Chinese is the equivalent of "limited liability corporation", again due to the phrase being on every single corporate web page.

The rest of the most popular 100,000 multi-grams fall in three categories. The first category is named entities, e.g. "University of California", the Chinese equivalent of "Shanghai Jiaotong University". The second category is many common speech patterns, e.g. "tell a friend", "if you can", "if I had". The third category is phrases, e.g. "fair market value", "real estate agents". Named entities appear to be the majority among the popular multi-grams, followed by phrases and then common speech patterns.

Thus, it's clear that frequency in a large corpora should be used as one of the signals for detection of entities and concepts, but should not be the only signal.

6. Related Work

There are many studies on using probabilistic methods to find significant N-grams, which corresponds to concepts and entities in human languages. The classic method is by Dunning [2], which uses the log-likelihood ratio to find significant bi-grams. Dunning's test has also been extended to detect N-grams for $N > 2$ [3].

Unlike the above approaches, this study does not attempt to find significant N-grams. Instead, this study simply examines the frequency distribution of all N-grams. To our knowledge, this study is the first to examine the similarities and differences between English and Chinese in terms of N-gram distributions using a large corpora ($>$ 200M web pages).

There are many approaches on using statistical methods to extract concepts and entities [4]. This study is very simplistic; only frequency information is used. However, manual examination of the top 100,000 N-grams suggest that simple signals such as frequency can be quite powerful at entity detections when the corpora is large enough.

7. Conclusions and Future Work

Statistics over a large corpora provide many interesting insights. In this study, we have shown that simple measures such as N-gram frequency distributions can illustrate differences and similarities between languages as different as English and Chinese. We have found that while 1-grams in Chinese are very different from 1-grams in English, multi-grams (3-grams and up) share nearly identical frequency distributions in English and Chinese.

We are in the process of extending our study to languages from different families, such as Russian and Arabic. We also plan to experiment with methods that detect significant N-grams and examine the frequency distributions of significant N-grams across different languages.

References

[1] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Proceedings of OSDI'04: Sixth Symposium on Operating System Design and Implementation*, 2004.

[2] T. E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[3] A. Franz and B. Milch. Searching the web by voice. *Proceedings of the 19th international conference on Computational linguistics*, 2:1–5, 2002.

[4] D. Lin and P. Pantel. Concept discovery from text.

[5] G. K. Zipf. Relative frequency as a determinant of phonetic change. *Reprinted from the Harvard Studies in Classical Philiology*, XL, 1929.