**ChineseWaC**

Siva Reddy

IIIT Hyderabad, India

[gvsreddy@students.iiit.ac.in](mailto:gvsreddy@students.iiit.ac.in)

Goal of this project is to collect Chinese web corpora of size two billion words (Two billion words might be little exaggerating).

**1. Brief Introduction to Chinese**

About one-fifth of the world’s population, or over one billion people, speak some form of Chinese as their native language. There are between seven and *thirteen* main regional groups of Chinese (depending on classification scheme), of which the most spoken, by far, are *Mandarin* (about 850 million), followed by *Wu* (90 million), *Cantonese* (Yue) (70 million) and *Min* (70 million). Some people call Chinese a language and its subdivisions as dialects, while others call Chinese a language family and its subdivisions as languages. The identification of the varieties of Chinese as "dialects" instead of "languages" is considered inappropriate by some linguists and Sinologists. This is mainly because most of these groups are mutually unintelligible (not understood without learning or much effort), although some, like *Xiang* and the *Southwest Mandarin* dialects, may share common terms and some degree of intelligibility.

**1.1 Chinese Writing System**

All the above varieties of Chinese have common writing systems. *Traditional Chinese* and *Simplified Chinese* characters are the two standard sets of printed Chinese characters (also called **hanzi**). Simplified Chinese character forms were created by decreasing the number of strokes and simplifying the forms of a sizable proportion of traditional Chinese characters. Simplified Chinese is mostly used in China (aka People Republic of China), Singapore and the United Nations whereas Traditional Chinese is used in Taiwan (aka Republic of China), Hong Kong and Macau. Debate on Simplified Chinese Vs Traditional Chinese can be found [here](http://en.wikipedia.org/wiki/Debate_on_traditional_and_simplified_Chinese_characters).

Web and electronic presence of Simplified Chinese is estimated to be more than Traditional Chinese.

*Pinyin* is currently the most commonly used romanization system for Chinese. Even this is used on web extensively (Mostly by foreign language learners and recent generations). Currently, standard Mandarin is taught in schools with Pinyin writing system.

**1.2 Chinese Morphology**

The number of Chinese characters contained in the [Kangxi dictionary](http://en.wikipedia.org/wiki/Kangxi_dictionary) is approximately 47,035, although a large number of these are rarely used variants accumulated throughout history. An educated Chinese uses about *three to four thousand* characters ( I was amazed to know this. OMG!!)

Chinese characters are *morphosyllabic,* each usually corresponding to a spoken syllable with a basic meaning. Chinese words are monosyllabic. However, a majority of words in Mandarin Chinese require two or more characters to write (thus are poly-syllabic) but have meaning that is distinct from the characters they are made from. So if we collect corpora of size two billion characters, it might be equal to one billion word corpora approximately.

Chinese has few grammatical inflections -- it possesses no tenses, no voices, no numbers (singular, plural; though there are plural markers, for example for personal pronouns), and only a few articles (i.e., equivalents to "the, a, an" in English). There is, however, a gender difference in the written language.

Chinese features Subject Verb Object (SVO) word order. Words are not separated by spaces which thus makes tokenization an important problem.

**1.3 Chinese Encodings**

Chinese character encodings can be used to represent text written in the CJK languages — Chinese, Japanese, Korean. Most common encodings used are

* Guobiao (prefixed by GB) mainly used by Simplified Chinese (PRC's official encoding)
* Big5 mainly used by Traditional Chinese (RC's official encoding)
* Unicode for all. (Accepted by every community. Famous recently)
* CP936 (GBK) and CP950 (Big5) [Microsoft's encodings]
* Other encodings are also present but not that popular.

Problem is the large character set reservation for these languages. Around 40000 characters are needed to represent the complete language. I am not going into those details.

Conversion from Traditional Chinese to Simplified is easy (many to one) but the reverse is quite difficult (one to many) and depends on the context. Most of them have been resolved now. Unicode to other conversions is very easy.

**2. Available Chinese Corpora**

Already existing corpora for Chinese are

**2.1 Chinese Gigaword Corpus**

It consists of comprehensive archive of Chinese news text acquired by LDC, University of Pennsylvania. (Corpus from 8 new's agencies). More details on website. I could not get the details of corpus size.

Website: <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T27>

**2.2 Chinese Internet and Business Corpus**

Sharroff (2006) collected web corpora using automated search engine queries. He used 500 frequent common words from the frequency list of Chinese Giga word corpus as seeds.

1. Chinese Internet Corpus, 280 million words (tokens). This corpus has been compiled by Serge Sharoff from the Internet in February 2005 along with other Internet corpora (for English, German and Russian).
2. Chinese Business Corpus, 30 million words (tokens). This corpus has been compiled by Serge Sharoff from the Internet in 2008 along with other business corpora (for English and Russian).

Website: <http://corpus.leeds.ac.uk/query-zh.html>

**2.3 Academia Sinica Balanced Corpus of Modern Chinese** (Mainly Traditional Chinese)

The Sinica corpus, a *Balanced Corpus* of Modern Chinese with 10 million words:

* 10 million words collected, primarily since 1996.
* Texts in the corpus are being collected from different areas and classified according to five criteria: genre, style, mode, topic, and source.
* Every text is segmented, and each segmented word is tagged with its pos.
* Represents the language used in Taiwan, and is therefore not representative of modern Mandarin Chinese as written on the mainland of China. [ [[1]](#footnote-0)Reference ]

Website:  <http://godel.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm>

<http://dbo.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh?language=1>

**2.4 Lancaster Corpus of Mandarin Chinese**

Created by Richard Xiao and Tony McEnery

* Balanced corpus of Mandarin Chinese for public use
* List of text categories present can be accessed [here](http://www.lancs.ac.uk/fass/projects/corpus/LCMC/lcmc/lcmc_kat.htm).

Website: <http://www.ling.lancs.ac.uk/corplang/lcmc/>

**3. Chinese Web as Corpus**

Here we aim to collect Chinese corpora using web.

**3.1 Earlier Work**

Emerson and O'Neil (2006) collected Chinese web corpora by crawling web mainly for the purpose of creating large lexicon. To create lexicon from text, one need to tokenize the text and to tokenize the text one need a lexicon. This becomes Chicken and Egg problem.

To create a large lexicon with completely unsupervised methods, we face the problem of precision. But Emerson and O'Niel argue that, with very large corpora, one can increase the precision at the expense of recall (Pasca 2004). As the corpus grows larger, it matters less what might be missed, since it will be seen again and again, but it becomes more important to avoid learning the noise and choking on the collected corpus data.

Emerson and O'Neil used Heritrix to crawl websites and explain the advantages of using Heritrix. Seed URL's are collected from Simplified Chinese section of the Open Directory Project (ODP). ODP is the most comprehensive human-edited directory of the Web. The ODP release dated 28 July 2005 was 210 MB compressed and 1 GB uncompressed, containing some 4.5 million URIs in 551,578 sections.

Many low level details which are very helpful to us are mentioned in the above paper. They were not able to remove near duplicates because of the size of the data they collected. But in our case, we already addressed this problem.

Other notable effort was by Sharroff (2006): He collected Chinese web corpora using automated search engine queries. Seed list used for generating queries can be accessed [here](http://corpus.leeds.ac.uk/internet/seeds-zh).

**3.2 Our approach**

I am planning to use the same approach used in our Corpus Factory. Seeds are collected from Chinese Wikipedia.

I think Bing can handle our load (It has certain usage principles which our programs abide by) and I support Honza statement that with more URL's more corpora of desired size can be downloaded. I will increase the number of queries to collect more URL's. I think we do not need to use a crawler.

Before I start it would be nice to take help of someone who have knowledge of Chinese tools. I could find many tools on web but I don't know which are the best tools available.

Resources:

* Chinese Wikipedia: I think it has no distinction between Simplified and Traditional Chinese. Need to confirm. Native Chinese people can confirm this. Targeted to the audience of both PRC and RC. Each variety of Chinese have a separate wiki resource.
* Need a word segmentation tool.
* Need a pos tagger.
* Need encoding converters.

Questions to be asked:

* What variety of Chinese do we plan to collect: Is it Mandarin, Wu or Cantonese or some other.
* What writing systems are we planning to include in our corpora: Simplified or Traditional or Pinyin (tonal + normal)

**References:**

Sharoff, S. (2006). “Creating general-purpose corpora using automated search engine queries”. In Marco Baroni and Silvia Bernardini, (eds), WaCky! Working papers on the Web as Corpus. Gedit, Bologna.Thomas Emerson and John O’Neil.  “Experience building a large corpus for Chinese lexicon construction,” in *In Marco Baroni and Silvia Bernardini, editors, WaCky! Working papers on the Web as Corpus* (Gedit, Bologna., 2006).

1. <http://www.lancs.ac.uk/fass/projects/corpus/LCMC/lcmc/lcmc_info.htm> [↑](#footnote-ref-0)