Chinese Web Scale Linguistic Datasets and Toolkit

Chi-Hsin Yu, Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University
#1, Sec.4, Roosevelt Road, Taipei, 10617 Taiwan
jsyu@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

ABSTRACT

The web provides a huge collection of web pages for researchers to study natural languages. However, processing web scale texts is not an easy task and needs many computational and linguistic resources. In this paper, we introduce two Chinese parts-of-speech tagged web-scale datasets and describe tools that make them easy to use for NLP applications. The first is a Chinese segmented and POS-tagged dataset, in which the materials are selected from the ClueWeb09 dataset. The second is a Chinese POS n-gram corpus extracted from the POS-tagged dataset. Tools to access the POS-tagged dataset and the POS n-gram corpus are presented. The two datasets will be released to the public along with their tools.

中文網路規模語言資料集和工具

網際網路提供研究人員巨量網頁進行自然語言處理研究，但是處理網路規模的文本不是件簡單的工作，而是需要大量的計算和語言資源。在本文中，我們介紹兩種加上中文詞性標記的網路規模資料集，以及易於將這些資源運用於自然語言處理應用的工具。第一種資料集選自於ClueWeb09的中文語料，並經過中文斷詞和詞性標記。第二種資料集是由上述詞性標記資料集中擷取中文詞性n-gram，所建立的語料庫。我們同時也提出搜尋詞性標記資料集和詞性n-gram語料庫的工具，這兩種資料集連同工具將提供研究人員使用。

KEYWORDS : Chinese POS n-gram, ClueWeb09, TOOLKIT
KEYWORDS IN L_2 : 中文詞性n-gram, ClueWeb09, 工具包

Proceedings of COLING 2012: Demonstration Papers, pages 501–508,
COLING 2012, Mumbai, December 2012.

1 Introduction

While using a large volume of data becomes a new paradigm in many NLP applications, preparing a web scale data collection are time consuming and need much cost. Recently, various versions of Gigawords which are comprehensive archive of newswire text data in Arabic, Chinese, English, French, and Spanish are distributed through LDC[1] to boost the researches. Besides newswire text, web n-grams in Chinese, English, Japanese, and 10 European languages created by Google researchers also released via LDC. Moreover, Lin et al. (2010) extend an English word n-gram corpus by adding parts of speech information and develop tools to accelerate the query speed.

In contrast to the web n-gram corpora, the ClueWeb[2] dataset developed by Carnegie Mellon University (CMU) contains a huge collection of raw web pages for researchers. It provides an alternative to construct corpora with fruitful context information rather than n-grams only. In this paper we extract the Chinese materials from the ClueWeb dataset, and develop two Chinese datasets, including a Chinese segmented and POS-tagged dataset called **PText**, and a Chinese POS n-gram corpus called **PNgram**. Besides, a toolkit is incorporated with these datasets.

This paper is organized as follows. Section 2 introduces the construction of the two Chinese corpora. Section 3 describes a Java-based tool for **PText**. Section 4 presents a user interface for the **PNgram**. Potential applications are also discussed in these two sections.

2 POS Tagging and N-gram Extraction

The ClueWeb09 dataset consists of about 1 billion web pages in ten languages. Based on the record counts, Chinese material is the second largest (177,489,357 pages) in ClueWeb09. Due to various charsets and encoding schemes, encoding detection, language identification and traditional Chinese-Simplified Chinese translation are indispensable preprocessing stages. Besides, enormous computational resources are needed for Chinese segmentation and part-of-speech tagging under this scale. The following sections show the procedures to construct two Chinese corpora and their basic statistics.

2.1 PText: A Chinese Segmented and POS-Tagged Dataset

Three tasks are described briefly as follows for the development of a Chinese segmented and POS-tagged dataset. The details can refer to Yu, Tang & Chen (2012).

(1) Encoding detection and language identification. Although the web pages in the ClueWeb09 dataset are encoded in UTF-8 encoding scheme, the correct encoding of a source page is still needed to be decided. In Chinese, web developers use many charsets and encodings to represent their web pages. For example, in traditional Chinese, there are charsets such as Big5, CNS 11643, and Unicode. In simplified Chinese, there are charsets such as GBK, GB2312, and Unicode. Furthermore, many ClueWeb09 web pages listed in Chinese category are actually in other languages such as Korean and Japanese. We must filter out those pages beforehand. Thus, encoding detection and language identification have

[1] http://www.ldc.upenn.edu/
[2] http://lemurproject.org/clueweb09.php/

to be done at the same time. Finally, 173,741,587 Chinese pages are extracted from the ClueWeb09 dataset.

(2) Chinese segmentation. A pure text in RFC3676 format is extracted from a web page for further linguistic processing. We translate all the web pages in traditional Chinese to simplified Chinese by using a character-based approach. Then, we split each Chinese web page into a sequence of sentences. The sentence boundaries are determined by full stop, question mark and exclamation mark in ASCII, full width and ideographic format. The new line characters '\r\n' are also used as a sentence boundary. Finally, we segment each sentence by using the Stanford segmenter (Tseng et al., 2005).

(3) Chinese POS tagging: The segmented sentences are tagged by using the Stanford tagger (Toutanova et al., 2003). The POS tag set of LDC Chinese Treebank is adopted. The tagger has been demonstrated to have the accuracy 94.13% on a combination of simplified Chinese and Hong Kong texts and 78.92% on unknown words.

The resulting dataset contains 9,598,430,559 POS-tagged sentences in 172,298,866 documents. In a document, we keep the original metadata such as title, the original TREC ID, and target URI in ClueWeb09. The encoding information in HTTP header and HTML header, and the detected encoding scheme are also preserved.

2.2 PNgram: A Chinese POS N-gram Corpus

For the extraction of the POS n-grams (n=1, …, 5), the first step is to determine the unigrams as our vocabulary. The minimum occurrence count of a unigram is set to 200, which is adopted in Chinese Web 5-gram (Liu, et al., 2010). After the unigram vocabulary is confirmed, the word *n*-grams (*n*=2, …, 5) are extracted. The minimum occurrence count of an n-gram (*n*=2, …, 5) is 40. After that, POS sequences for each word *n*-gram are extracted. The dataset format is like Google N-gram V2 (Lin et al., 2010). The following shows examples with their English translation.

唸书 时间 VV NN 2 | NN NN 119
(read, time)
唸书 期间 NN NN 43
(read, period)
唸书 时期 VV NN 2 | NN NN 73
(read, period)

Table 1 shows word n-gram statistics of the resulting PNgram corpus. There are 107,902,213 unique words in PText, and only 2.1% of unique words (2,219,170) with frequency larger than 200. For a word bigram, the minimum frequency is 40. Total 9.7% unique bigrams are selected. The ratio decreases roughly half when N increases.

n-gram	#PText entries	# PNgram entries	Ratio
1	107,902,213	2,219,170	2.1%
2	645,952,974	62,728,971	9.7%
3	4,184,637,707	200,066,527	4.8%
4	10,923,797,159	294,016,661	2.7%
5	17,098,062,929	274,863,248	1.6%

TABLE 1 –Statistics of the word N-grams

Table 2 shows the statistics of the extracted PNgram corpus. The Stanford POS tagger (Toutanova et al. 2003) adopts LDC Chinese Treebank POS tag set in the trained Chinese tagger model. There is 35 POS tags plus one additional tag STM for sentence start mark <S> and stop mark </S>. We can see that the average POS patterns per word patterns are small. In other words, they range from 1.7 to 3.5 POS patterns.

n-gram	Avg. POS patterns per word n-gram	Max. POS patterns of word n-gram
1	3.5	20
2	2.5	301
3	2.3	2,409
4	2.1	9,868
5	1.7	7,643

TABLE 2 –Statistics of POS patterns in **PNgram**

3 Tools for the PText Dataset

The PText dataset are stored in Java serialization format which is easy to be manipulated by programmers.

3.1 Data Model and Tools

We briefly describe the data model and tools for the PText dataset as follows.

(1) Data model. The dataset contains 4,395 gzipped files. Each gzipped file contains a list of document objects, where a document contains a list of sentences, and a sentence contains a list of tagged words. In this way, it is very easy to traverse the whole dataset. There is no need to re-parse the data from plain texts.

(2) Tools. Java classes of data model are provided along with Java classes that are used to de-serialize the gzipped files. It is easy to traverse the whole dataset in parallel by programmers. Conversion tools are also provided for dataset users who want to convert the gzipped files to readable plain texts.

3.2 Applications with the PText dataset

The PText dataset is beneficial for many potential researches. We outline some interesting applications below.

(1) Knowledge mining. The PText dataset can be used as the source of OpenIE (Etzioni et al., 2008). In ReVerb, Fader, Soderland, and Etzioni (2011) use simple POS patterns to mine facts from the web texts in English. Similarly, researchers can extract facts from Chinese texts by specifying Chinese POS patterns. With the PText dataset, considerable pre-processing time can be saved for Chinese OpenIE researchers.

(2) Sentiment analysis. As shown by the papers (Wiebe et al., 2004; Abbasi, Chen, and Salem, 2008), parts-of-speech information is useful for many sentiment analysis tasks such as opinion classification and subjectivity analysis of web texts. With the large-scale PText dataset, researchers can investigate more rich phenomena in the web texts.

(3) Basic NLP tasks. Besides the new and interesting researches, the fundamental NLP tasks can also benefit from the PText dataset, e.g., the encoding detection and language identification in pre-processing Chinese web pages, the performance of Chinese word segmenters and POS taggers in large scale web texts, and so on.

4 A User Interface for the PNgram Corpus

Lin et al. (2010) provide source codes to search English POS n-gram data along with a Lisp-style query language. The query tool is powerful, but it is not intuitive for users of non-computer background. In this paper, we design an easy-to-use interface for users.

4.1 User Interface

Figure 1 shows the snapshot of the user interface to access the PNgram Corpus. At first, users input the length of the requested n-gram and the range of its frequency. By default, the minimum frequency is 40. Then system will provide suitable number of slots for users to write down the linguistic feature of each word. Users can fill in a wildcard * or a word, along with its lexical information specified in the constraint part. The possible constraints include a duplication form like AA for searching a duplication word 哈哈 (ha ha), and possible POS tags for filtering non-relevant patterns.

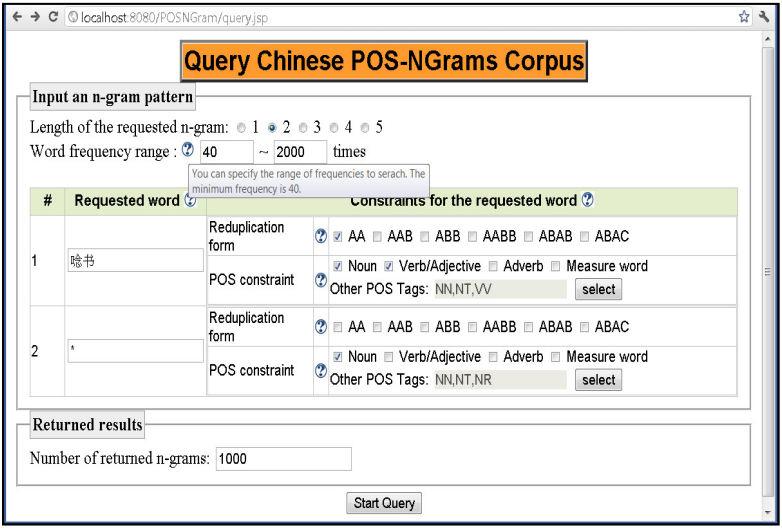

FIGURE 1 –A user interface to query the Chinese POS-NGram corpus

In Figure 1, users search bigram patterns. The first word is 唸书 (read), and it must be Noun, Verb/Adjective, or tags NN, NT, VV, which are selected in Figure 2. The second word can be any words with Noun, or NN, NT and NT tags. The word frequency is limited between 40 to 2,000 times. The returned POS n-gram results are similar to the examples shown in Section 2.2.

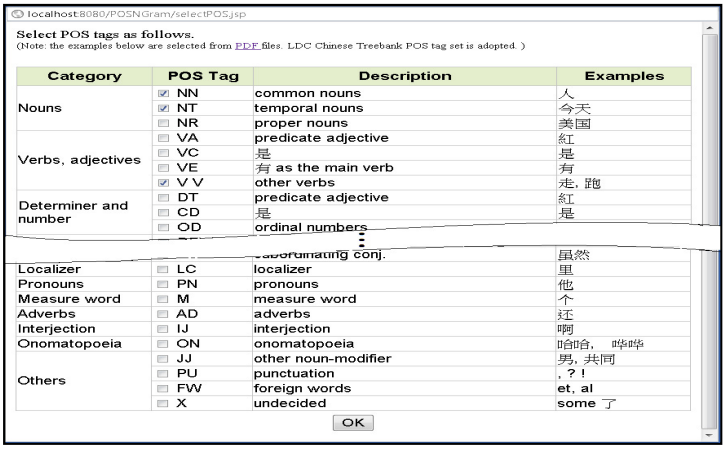

FIGURE 2 –Select specific POS tags for a word

4.2 Applications of the PNgram Corpus

The PNgram corpus along with the user interface is useful for many NLP applications. Yu and Chen (2012) employ it to detect Chinese word order errors. We outline some others as follows.

(1) Language Learning. In Chinese learning, we may be interested in what linguistic context some specific reduplication forms like "快快樂樂" (happy happy) appear. Through the interface, it is easy to collect their usages from the web data. Similarly, the uses of measure words in Chinese sentences are very common. The PNgram corpus along with the tool provides a flexible way to analyze the measure words for a specific word in web texts.

(2) Pattern Identification for Information Extraction (IE). In IE, a named entity usually ends with some specific characters such as 站/station in 雷达站/radar-station. But the character 站/station can be a verb in word 站在/stand-in. The PNgram corpus and the accompanying tool can be used to collect similar NE patterns satisfying the POS criterion for NE rule extraction.

(3) Chinese Noun Compound Corpus Construction. As a concept is usually represented by a multiword expression, the structure of a noun compound is needed to be determined. We can specify a POS pattern to construct a noun compound corpus from the PNgram dataset, and use it to study the modifying structures of noun compounds.

5 Conclusion and Future Work

In this paper, we present the POS tagged dataset (**PText**) and the POS 5-gram corpus (**PNgram**). Besides, we provide tools for users to access these two resources. Researchers can collect sentences from the web-scale linguistic resources for their own specific research topics. For example, we will study the polarity of Chinese discourse markers based on these web-scale corpora. Sentences where discourse markers occur will be extracted, sentiment polarities of the discourse arguments connected by a discourse marker will be measured, and the relations between discourse parsing and sentiment analysis will be investigated.

Acknowledgments

This research was partially supported by Excellent Research Projects of National Taiwan University under contract 101R890858 and 2012 Google Research Award.

References

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems*, 26(3), 12:1–12:34.

Brants, T., and Franz, A. (2006). Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open Information Extraction from the Web. *Communications of the ACM*, 51(12):68–74.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying Relations for Open Information Extraction. In the *Conference on Empirical Methods in Natural Language Processing* (*EMNLP '11*), pages 1535–1545, Edinburgh, Scotland, UK.

Gao, J., Goodman, J., Li, M., and Lee, K.-F. (2002). Toward a Unified Approach to Statistical Language Modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, 1(1):3–33.

Lin, D., Church, K., Ji, H., Sekine, S., Yarowsky, D., Bergsma, S., Patil, K., Pitler, E., Lathbury, R., Rao, V., Dalwani, K. and Narsale, S. (2010). New Tools for Web-Scale N-grams. In *the Seventh conference on International Language Resources and Evaluation*, pages 2221–2227, Malta.

Liu, F., Yang, M., and Lin, D. (2010). Chinese Web 5-gram Version 1. Linguistic Data Consortium, Philadelphia.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Edmonton, Canada.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In the *Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171, Jeju, Korea.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning Subjective Language. *Computational Linguistics*, 30(3):277–308.

Yu, C.-H., Tang, Y. and Chen, H.-H. (2012). Development of a Web-Scale Chinese Word N-gram Corpus with Parts of Speech Information. In *the Eighth International Conference on Language Resources and Evaluation*, pages 320–324, Istanbul, Turkey.

Yu, C.-H. and Chen, H.-H. (2012). Detecting Word Ordering Errors in Chinese Sentences for Learning Chinese as a Foreign Language. In *the 24th International Conference on Computational Linguistics*, Mumbai, India.