Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus

Wei-Yun Ma, Chu-Ren Huang
Academia Sinica
128 Sec. 2, Academia Rd, Nankang, Taipei 115 Taiwan, R.O.C.
ma@iis.sinica.edu.tw, churen@sinica.edu.tw

Abstract

Tagging as the most crucial annotation of language resources can still be challenging when the corpus size is big and when the corpus data is not homogeneous. The Chinese Gigaword Corpus is confounded by both challenges. The corpus contains roughly 1.12 billion Chinese characters from two heterogeneous sources: respective news in Taiwan and in Mainland China. In other words, in addition to its size, the data also contains two variants of Chinese that are known to exhibit substantial linguistic differences. We utilize Chinese Sketch Engine as the corpus query tool, by which grammar behaviours of the two heterogeneous resources could be captured and displayed in a unified web interface. In this paper, we report our answer to the two challenges to effectively tag this large-scale corpus. The evaluation result shows our mechanism of tagging maintains high annotation quality.

1. Background

With growing interest in Chinese language processing, a few gargantuan Chinese corpora of modern Chinese have been assembled and released with query tools in recent years. For example, the Sinica Corpus (CKIP, 1995/1998) developed by Academia Sinica in Taiwan contains 5.2 million words with part-of-speech tag (POS) while the Chinese corpus developed by the Center for Chinese Linguistics (CCL corpus) at Peking University contains 85 million Chinese characters. Both corpora offer the keyword-in- context (KWIC) function for inspecting the context of a given keyword through their web interfaces. However, there are two major restrictions to use the both popular online corpora to obtain deeper and comparable Chinese grammatical information. One restriction is that although the Sinica Corpus is segmented and POS-tagged, CCL is not segmented and tagged. Therefore it is unable to make deeper syntactic analysis via CCL and is also difficult to compare the syntactic behaviours of a given word between Taiwan and Mainland China. The other difficulty is that only utilizing KWIC concordance is not sufficient to capture and display complete and organized grammatical information of a given keyword.

Other several existing linguistic annotated corpora of Chinese, e.g. Penn Chinese Tree Bank (Xia *et al* 2000, Xue *et al* 2002), Sinica Treebank (Huang *et al* 2000), provide more elaborate annotations. But they suffer from the same problem: they are all extremely labor-intensive to build and typically have a narrow coverage and are therefore insufficient to reflect the real usage of a given keyword.

In this paper, in order to resolve the difficulties above, we attempt to segment and POS-tag the Chinese Gigaword Corpus (CGW) released in 2003 by Linguistic Data Consortium (LDC). CGW was produced by LDC. It contains about 1.12 billion Chinese characters, including 735 million characters from Taiwan's Central News Agency (CNA) from 1991 to 2002, and 380 million characters from Mainland China's Xinhua News Agency (XIN) from 1990 to 2002. CNA uses the complex character form and XIN uses the simplified character form. CGW has three major advantages for the corpus-based Chinese linguistic research: (1) It is large enough to reflect the real written language usage in either Taiwan or Mainland China. (2) All text data are presented in a SGML form, using a markup structure to provide each document with rich metadata for further inspecting. (3) CGW is appropriate for the comparison of the Chinese usage between Taiwan and Mainland China, because it provides the same newswire text type, and these news texts were almost published during the overlapping time period. We utilize Chinese Sketch Engine (Kilgarriff *et al* 2004, Kilgarriff *et al* 2005) as the corpus query tool, by which grammar behaviours of the two heterogeneous resources could be captured and displayed in a unified web interface. Therefore, how to annotate the two heterogeneous corpora to let them could be consistently compared their words' syntactic behaviours through Chinese Sketch Engine is an important concern in this paper.

A challenging task is to segment and POS-tag such huge amount of corpus efficiently. Given the corpus size, it is clearly not possible to adopt the semi-automatic approach of human-aided machine tagging to reach the task in the limited time. Therefore, even adopting full-automatic tagging strategy but still maintaining high annotation quality is also a major task in this paper.

2. Introduction to CGW

We begin with an introduction to the details of CGW because of its importance at our processing strategies.

2.1. Size of CGW

Table 1 presents the following categories of information: source of the data, number of files per source, Totl-MB shows totals for file sizes (nearly 4 gigabytes, total), number of characters, and number of documents.

Source	#Files	Totl-MB	K-#Chars	#DOCs
CNA	144	2606	735499	1649492
XIN	142	1331	382881	817348
TOTAL	286	3937	1118380	2466840

Table1. Size of CGW

Taiwan's CNA is from 1991 to 2002, and Mainland China's XIN is from 1990 to 2002. Each file contains all documents for the given month from the given news source.

2.2. SGML Form

All text data are presented in a SGML form, using a very simple, minimal markup structure. The markup structure, common to all data files, can be illustrated by the following example:

```
<DOC id="CNA19910101.0003" type="story">
<HEADLINE>
捷運局對工程噪音採多項防治措施
</HEADLINE>
<DATELINE>
(中央社台北一日電)
</DATELINE>
<TEXT>
<P>
台北都會區捷運工程正處於積極趕工階段,…
</P>
<P>
淡水線工程進度百分之三十六點一九,落後百分之二點六七,…
</P>
</TEXT>
</DOC>
```

Figure1. Example of a news document in CGW

For every "opening" tag (DOC, HEADLINE, DATELINE, TEXT, P), there is a corresponding "closing" tag. The "id=" attribute of DOC consists of the 3-letter source abbreviation (in CAPS), an 8-digit date string representing the date of the story (YYYYMMDD), a period, and a 4-digit sequence number starting at "0001" for each date (e.g. " CNA19910101.0003"); in this way, every DOC in the corpus is uniquely identifiable by the id string.

3. Design of Automatic Annotator

There are two major missions of our automatic annotator: word segmentation and POS tagging. In order to speed up the process and to maintain high quality at the same time, our automatic annotator has the following characteristics: (1) The annotator takes advantage of the characteristics of CGW for reaching high annotation quality. (2) The annotator has the capability to process a large corpus efficiently, which means the program is robust, and hardware resources used by the program are carefully managed. (3) The annotation format exerts the merits of the used corpus query tool (i.e. Chinese Sketch Engine). (4) The annotator generates some records of annotation process for speeding up human examination if human examination is still decided to be done in the future. For instance, several word types are more difficult to be correctly identified. The annotator records the list of these unreliable words. If human examination is undertaken in the future, human annotators will only need to examine these records and get much better whole quality in a limited time.

We enhanced Sinica Word Segmenter (Ma and Chen 2005) to possess the above characteristics. And we utilized HMM method for POS tagging and morpheme-analysis-based method (Tseng and Chen 2002) to predict POSs for new words.

3.1. Document-based v.s. Corpus-based Statistical Information

Occurrences of new words, which are not covered in the lexicon, degraded significantly the performances of most word segmentation methods. The number is especially higher in news reports-averagely 3% to 5% new words within a news document. Therefore unknown word identification would play a key role for segmenting CGW.

Most popular segmentation technologies (Chiang 1992, Tseng 2005) use corpus-based statistical methods for identifying new words with high statistics and use morphological rules for those with low statistics. However, for these corpus-based statistical methods, they usually suffer a problem that phrases or partial phrases are easily incorrectly identified as words because of their statistical significance in a corpus. Even very frequently superfluous character strings with strong statistical associations are also easily incorrectly identified as words. Similarly, on the other side, frequently new words with high statistics within a document are probably hard to be identified because of their low statistics in a whole corpus. This situation is more serious while processing newswire text data. For newswire text data like CGW, a document usually tightly focuses on the same event or subject, and the keywords of a text are often new words and frequently recur in a news document, but not necessarily recur the same proportion in the whole corpus.

Therefore, for statistical methods of our word segmentation, we mainly rely on the document-based statistical information instead of corpus-based statistical information so that the locality of the keywords in a newswire document is fully utilized. Because all text data of CGW are presented in a SGML form, it is convenient to separate CGW into individual documents using a simple SGML parser. We proposed two strategies of word segmentation by pseudocodes shown in Figure 2 and Figure 3.

In Strategy A, while segmenting a given document, only the basic lexicon and extracted new words of the document are referenced. In Strategy B, while segmenting a given document, we also references NewWordLexicon collected from other documents. But two things are worth noticing: One is that in NewWordLexicon only new words with high accumulated frequency are covered, which means these words have high reliability as real words. Another is that when referencing these statistics, the statistics of a given document should still play a more important role than NewWordLexicon for resolving segmentation ambiguity.

In addition to fully utilizing locality of newswire data text, Strategy A or B also has another advantage: the memory resource is always controlled within the range of a document, which also means the total processing time will be much shorter than corpus-based statistical methods because the searching space of document-based statistical information is much smaller than the searching space of corpus-based statistical information.

For each newswire document- d_i
Begin
  Calculate statistical information- s_i from d_i
  Extract out new words- nw_i by referencing s_i and (probabilistic) morphological rules
  Segment d_i by referencing the basic lexicon and nw_i
  Release memory resources for d_i, s_i, nw_i
End

Figure2. Strategy A

For each newswire document- d_i
Begin
Calculate statistical information - s_i from d_i
Extract out new words- nw_i by referencing s_i and (probabilistic) morphological rules
Release memory resources for d_i, s_i, but keep the record of nw_i
End
For each new word in the collection of all nw, accumulate its frequency from the records of all nw and collect those new words which accumulated frequencies are higher than a threshold. The filtered collection is named as NewWordLexicon
For each newswire document- d_i
Begin
Segment d_i by referencing the basic lexicon, nw_i, and NewWordLexicon
Release memory resources for d_i, nw_i
End

Figure3. Strategy B

3.2. Annotation Format

We utilize Chinese Sketch Engine as the corpus query tool. Besides traditional KWIC function, the engine would automatically generate a one-page, corpus-derived summary of a given word's grammatical and collocation behaviour, such as the distributions of its subjects, objects, preposition objects, and modifiers, by consulting grammatical relations for Chinese. The grammatical relations are defined using regular expressions over POS tags. The more elaborate grammar relations are, the more precise querying results will be obtained.

Therefore in order to facilitate the design of flexible and elaborate grammar relations of Chinese Sketch Engine, we adopted mixing POSs tagging strategy: after segmentation and HMM-based tagging process, each word is annotated with the basic POS, such as "陳(Nb[1])". And for most words, their basic POSs can be further converted into elaborate POSs, such as "陳(Nbc[2])", by consulting the basic lexicon. The rest of words, such as new words, are still reserved with their basic POSs, which are obtained by the prediction of the morpheme-analysis-based tagger. The final annotation results can be illustrated by the Figure 4 (Bold characters represent new words and their predicted basic POSs, the others represent words and their elaborate POSs covered in the basic lexicon, or quantifier words, reduplicated words, etc):

<DOC id="CNA19910101.0003" type="story">
<HEADLINE>
捷運局(Nc) 對(P31) 工程(Nac) 噪音(Nad) 採(VC2) 多(Neqa) 項(Nfa) 防治(VC2) 措施(Nac)
</HEADLINE>
<DATELINE>
((PARENTHESISCATEGORY) 中央社(Nca) 台北(Nca) 一日(Nd) 電(VC2))(PARENTHESISCATEGORY)
</DATELINE>
<TEXT>
<P>
台北(Nca) 都會區(Ncb) 捷運(Nad) 工程(Nac) 正(Dd) 處於(VJ3) 積極(VH11) 趕工(VA4) 階段(Nac)
，(COMMACATEGORY) …
</P>
<P>
淡水線(Na) 工程(Nac) 進度(Nad) 百分之三十六點一九(Neqa)
，(COMMACATEGORY)
落後(VJ1) 百分之二點六七(Neqa) ，(COMMACATEGORY)…
</P>
</TEXT>
</DOC>

Figure4. An annotation example

4. Implementation

In order to exhibit substantial linguistic differences under consistent querying environment for CNA and XIN, it is necessary to use a unified basic lexicon and POS tagset for annotation. The basic lexicon we used consists of three sources: (1) Sinica lexicon with 80000 word entries. (2) A 50000-words' set collected from Sinica Corpus 3.0, which is a balanced corpus of modern Chinese containing separated words and their POSs checked by human. (3) Xinhua new-words lexicon, which collects 5000 new words frequently used in Mainland China. We adopt Sinica Tagset as the uniform POS tagset for CNA and XIN.

So far we have finished implementing Strategy A discussed in sention 3.1. An array of machine was used to process CGW, which took over 3 days to perform. After completing the whole annotation of CGW, total 462 million words of CNA and 252 million words of XIN are identified.

4.1. Evaluation

We randomly picked one document from CNA per season and one document from XIN per year. Then there are total 48 documents of CNA and 12 documents of XIN. They are regarded as testing data set for evaluation. These 60 documents are carefully checked by a linguist. The annotation performance is provided in Table 2.

	RefWord#	TestWord#	MatchWord#	Recall	Precision
CNA	12416	12500	12186	0.98	0.97
XIN	3945	4002	3790	0.96	0.95

Note: Recall=MatchWord# / RefWord#
Precision=MatchWord# / TestWord#

	MatchWord#	MatchPOS#	POS Precision
CNA	12186	12033	0.99
XIN	3790	3725	0.98

Note: POS Precision= MatchPOS# / MatchWord#

Table2. Evaluation result

The evaluation result shows that our automatic annotater performs very well in either CNA or XIN. The segmentation performance of XIN is a bit lower than CNA probably because most of the words in our basic lexicon are collected from Taiwan sources. In other words, the proportion of new words of XIN are higher than CNA, and these new words caused rather more segmentation mistakes.

[1] "Nb" represents "proper noun" according to Sinica Tagset.
[2] "Nbc" represents "Chinese surname", one kind of proper noun, according to Sinica Tagset.

4.2. Character Form Conversion

To clearly and conveniently observe querying results of a given word appeared in CNA and XIN, the distinct character forms need to be unified as the same as a given querying word's form. Therefore we in advance generated two additional data sets: CNA with the simplified character form obtained through the conversion of its original complex character form, and XIN with the complex character form obtained through the conversion of its original simplified character form. Therefore four data sets were obtained. We further generated another two data sets through combining the existed four data sets: one data set is generated through combining CNA and XIN with the complex character form, the other data set is generated through combining CNA and XIN with the simplified character form. Word Sketch Engine then could directly display the querying results of CNA and XIN with the same character form at the same time. The examples are shown as Figure 5 and Figure 6.

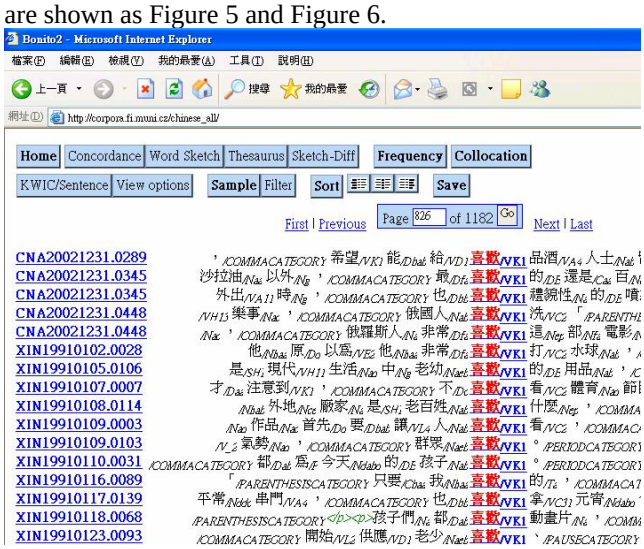

Figure5. KWIC concordance result with the complex character form while querying word "喜歡" of the complex character form.

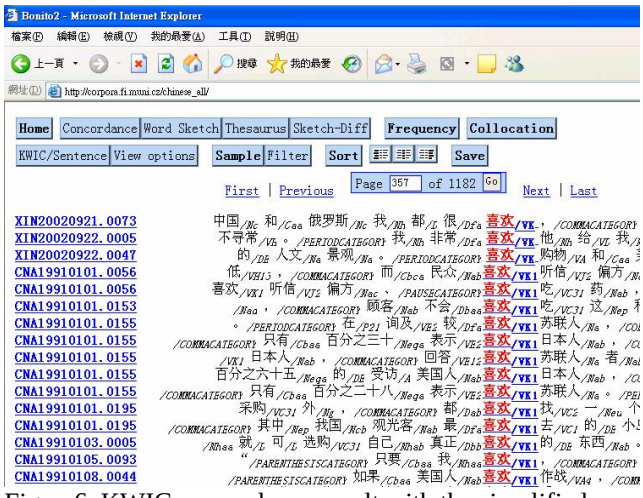

Figure6. KWIC concordance result with the simplified character form while querying word "喜欢" of the simplified character form.

5. Conclusion and Future Work

Based on careful analyses of CGW's characteristics, in this paper, we proposed our concerns and strategies for tagging CGW. Not much processing time and high annotation quality demonstrate our automatic annotator performs very well. We also concerned about the relation between the corpus query tool and the annotated corpus. How to fully exert the advantages of the corpus query tool is an important concern about the design of the annotation strategy and the annotation format. In our work, we utilized the same lexicon and tagset to segment CNA and XIN, by which Word Sketch Engine could exhibit substantial linguistic differences under consistent querying environment of the heterogeneous sources-respective news in Taiwan and in Mainland China.

We are now collecting more lexicon resources from Mainland China in order to further improve the segmentation performance of XIN in the future. We are also working on another related project-to automatically mark nominalization feathers on those verbs in CGW with noun usages in specific contexts.

We expect our experiences of tagging CGW will be a worthy example to reference for the development of any gargantuan and heterogeneous corpus.

6. References

CKIP (Chinese Knowledge Information Processing Group). (1995/1998). The Content and Illustration of Academica Sinica Corpus. *(Technical Report no 95-02/98-04). Taipei: Academia Sinica*

Xia, Fei, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. (2000). Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. *Proceedings of LREC*

Xue Nianwen, Fu-Dong Chiou, and Martha Palmer (2002). Building a Large-Scale Annotated Chinese Corpus *Proceedings of COLING*

Huang Chu-Ren, Keh-Jiann Chen, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao and Kuang-Yu Chen. (2000). Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface. *Proceedings of 2nd Chinese Language Processing Workshop* pp. 29-37.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell. (2004). The Sketch Engine. *Proceedings of EURALEX*, Lorient, France.

Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychlý, Simon Smith, and David Tugwell. (2005). Chinese Word Sketches. *ASIALEX 2005: Words in Asian Cultural Context.*

Ma, Wei-Yun and Keh-Jiann Chen, (2005). Design of CKIP Chinese Word Segmentation System, *Chinese and Oriental Languages Information Processing Society*, Vol 14. No. 3. pp. 235-249.

Tseng, H.H. & K.J. Chen, (2002). Design of Chinese Morphological Analyzer," *Proceedings of SIGHAN Workshop on Chinese Language Processing*, pp. 49-55

Chiang, T. H., M. Y. Lin, & K. Y. Su, (1992). Statistical Models for Word Segmentation and Unknown Word Resolution, *Proceedings of ROCLING V*, pp. 121-146.

Tseng, H.H., Pichuan Chang, Galen Andrew, Daniel Jurafsky, Christopher Manning. (2005). A Conditional Random Field Word Segmenter. *Proceedings of SIGHAN Workshop on Chinese Language Processing.*