Quality Assurance of Automatic Annotation of Very Large Corpora: a Study based on Heterogeneous Tagging Systems

Chu-Ren Huang[1], Lung-Hao Lee[1], Wei-guang Qu[2], Jia-Fei Hong[1], Shiwen Yu[2]
[1]Institute of Linguistics, Academia Sinica
[2]Institute of Computational Linguistics, Peking University
[1]128 Sec. 2, Academia Rd., Nankang, Taipei 115 Taiwan, R.O.C.
[2]Beijing, 100871, China
churen@gate.sinica.edu.tw, lunghao@gate.sinica.edu.tw, wggu@pku.edu.tw,
jiafei@gate.sinica.edu.tw, yusw@pku.edu.tw

Abstract

We propose a set of heuristics for improving annotation quality of very large corpora efficiently. The Xinhua News portion of the Chinese Gigaword Corpus was tagged independently with both the Peking University ICL tagset and the Academia Sinica CKIP tagset. The corpus-based POS tags mapping will serve as the basis of the possible contrast in grammatical systems between PRC and Taiwan. And it can serve as the basic model for mapping between the CKIP and ICL tagging systems for any data.

1. Motivation and Goals

Quality assurance of automatically tagged corpora has become a central issue in the study of language resources. As very large corpora, such as those constructed from the web-as-corpus approach (Kilgarriff & Grefenstette, 2003) become the norm, it also becomes obvious that manual checking of tagging and other textual markup will not be feasible. It is essential that automatic quality assurance measures can be devised. Previous research on quality assurance of POS tagging presupposed one tagset and try to discover distributional anomalies of word-tag pairs. In an ideal situation, two different automatic taggers can be employed and the inconsistencies in their results will be resolved to improve both precision and recall. However, as automatic tagging techniques become more optimized and harmonize, such inconsistencies became rarer, yet the shared mistakes become even harder to cover. Our current study takes a different assumption. Suppose there are two competing tagsets (presumably two similar but different linguistic analysis systems) available, a substantial number of discrepancies can be expected. Comparison of two versions of the same corpus allow discovery of both regular mapping and non-regular mappings. Non-regular mapping can be further analyzed to identify both potential errors and systematic correspondences.

This two tagset model is a viable alternative when a language has more than one commonly accepted tagsets, such as in English. It is even necessary when a language contains significant variants, such as in Mandarin Chinese. In Mandarin Chinese, the PRC and Taiwan as developed two significant variants. It is well-established even among Chinese computational linguists that PRC corpora are best processed with PRC standards, and vice versa. However, there are obvious motivations for uniform markup of both variants, such as for web-based information retrieval and for corpus-based comparative studies of the two variants. The LDC Chinese GigaWord Corpus is designed with such cross-variation research purposes in mind. When such a heterogeneous corpus is tagged, there are two competing requirements for tagging. First, a uniform tagset for all data is desirable for study of linguistic generalizations, both shared and contrastive between two variants. On the other hand, the linguistic system of each variant is best represented with its locally accepted tagset.

In this paper, we propose a set of heuristics for improving annotation quality in such huge amount of corpus efficiently. The Xinhua News portion of the Chinese Gigaword Corpus was tagged independently with both the Peking University ICL tagset and the Academia Sinica CKIP tagset. The ICL tagged portion was automatically tagged without proofreading, while the CKIP tagged corpus was previously checked. By comparing these two different versions of tagged the same corpus, we hope two goals: the first is to devise a (semi-)automatic way of error-detection for quality assurance of the fully automatically tagged ICL version of the corpus. The other is to establish empirically attested mapping between the two tagsets. The corpus-based mapping, annotated with probability of mapping relations, will serve as basic data for two very different purposes. First, it will serve as the basis of the possible contrast in grammatical systems between PRC and Taiwan. Second, it can serve as the basic model for mapping between the CKIP and ICL tagging systems for any data.

2. Background to Chinese Gigaword

Automatic POS annotation is remains a challenging task in Chinese language processing. For instance, ACL SigHan has hosted four bakeoff competition for segmentation, but non for POS tagging. There is only a handful of POS tagging systems and automatic taggers which are widely accepted and accessible. In Taiwan, Academia Sinica’s CKIP tagset has been considered the standard and has been used in annotating the Sinica Corpus (CKIP, 1995/1998), which were first annotated in 2006 and contains roughly 10 million words in the latest version (2007). In PRC, the Institute of Computational

Linguistics (ICL)’ s tagset has been considered the de facto standard and is widely available through the POS tagged People’s Daily Corpus. However, an even greater challenge occurs with the new demand of very large corpora and the availability of the untagged LDC Gigaword Corpus.

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.

SGML form, a very simple and minimal markup structure originally designed by LDC, can be illustrated by the following example. 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), and a 4-digit sequence number string at “0001” for each date. For example, the id attribute named as “XIN20010101.0004” is uniquely identifiable to the DOC in the corpus.

```
<DOC id="XIN_CMN_20010101.0004" type="story">
<HEADLINE>
印度平静迎接新千年
</HEADLINE>
<DATELINE>
新华社新德里 1 月 1 日电
</DATELINE>
<TEXT>
<P>
(记者熊昌义)10 亿印度人以平静的心情迎来了新千年。虽然
官方 1 日没有计划举行迎接新千年的活动,但印度首都新德
里一些著名的五星饭店、购物中心,商业街和集贸市场,到处
张灯结彩,火树银花,充满了节日气氛。
</P>
<P>
1 日凌晨,新德里虽下起了小雨,但马路上仍然有不少车辆在
行驶,街头可以看到一群群青年人伴随着欢乐的乐曲翩翩起
舞。偶而还能看到人们施放的礼花,听到一些鞭炮声。
</P>
........
</TEXT>
</DOC>
```

Figure 1: An example of news document in CGW

3. Resources: Chinese Gigaword Corpus Tagged with two Different Tagsets

3.1 Academia Sinica’s CKIP Annotator

There are two major missions of CKIP automatic annotator: word segmentation and POS tagging. We enhanced Sinica Word Segmenter (Ma & Chen, 2005) to segment the corpus into the words. And we utilized HMM method for POS tagging and morpheme-analysis-based method (Tseng & Chen, 2002) to predict POSs for new words. All the annotated text is traditional characters in Big5 encoding. And the full numbers are adopted. Fig. 2 shows an example with CKIP-POS tags.

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.

LDC’s Chinese Gigaword Corpus currently has a segmented and tagged version available. This version adopts the CKIP tagset and was performed automatically with automatic and partially manual post-checking (Ma & Huang, 2006). The precision accuracy is estimated to be over 95% for Central New Agency part of data from Taiwan. However, for the Xinhua New Agency data from PRC, they were not able to independently verify their accuracy. In addition, it would be very helpful to have the PRC data tagged with the ICL-PKU tagset such that they can be easily compared to existing literature and also be accessible to other NLP applications developed in PRC.

3.2 Specification for ICL-PKU Tagset

The institute of Computational Linguistics, Peking University made a specification (as known as Specification 2001) for the word segmentation and POS tagging of its People Daily corpus (Yu et al, 2002). The size of this corpus is over 26 million Chinese characters. In order to build the phonetically annotated corpus (1 million Chinese characters), the added Phonetic Notation was made in Specification 2003 (Yu et al., 2003).

A team from PRC applied a machine-learning based algorithm to automatically tag the Xinhua data. They presented a unified approach for Chinese lexical analysis using Hierarchical Hidden Markov Model (HMMM), which named as ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) aiming to incorporate Chinese word segmentation, POS tagging, disambiguation and unknown words recognition into a whole theoretical frame (Zhang et al., 2003a, b).

The result of tagging Xinhua data was completed without human intervention. All the tagged text was simplified characters in GBK encoding. This automatically tagged data is then compared with the CKIP tagged portion of the same corpus for both comparison and quality assurance purposes. Fig. 3 shows an example with

PKU-POS tags.

3.3 Preprocessing on POS Tagged Corpus

Given the same representational format, the following preprocessing procedures were taken to eliminate encoding inconsistencies and character variations:

Steps 1: All the simplified characters were converted into traditional characters.

Steps 2: Language encoding of all characters were converted into Unicode (UTF-8)

Steps 3: All full numbers in text are converted into common numbers for consistency.

```
<DOC id="XIN_CMN_20010101.0004" type="story">
<HEADLINE>
印度(Nca)  平靜(VH11)  迎接(VC2)  新(VH11)  千(Neu)
年(Nfg)
</HEADLINE>
<DATELINE>
新華社(Nca)  新德里(Nca)  １月(Nd)  １日(Nd)  電(Naa)
</DATELINE>
<TEXT>
<P>
((PARENTHESISCATEGORY)  記 者 (Nab)  熊 昌 義
(Nb)  )(PARENTHESISCATEGORY)  １０億(Neu)  印度
人(Nab) 以(P11) 平靜(VH11) 的(DE) 心情(Nad) 迎(VC2)
來(VA11)  了(Di)  新(VH11)  千(Neu)  年(Nfg)  。
(PERIODCATEGORY) ........
</P>
<P>
１日(Nd)  凌晨(Ndabe)  ，(COMMACATEGORY)
新德里(Nca)  雖(Cbba)  下(VC)  起(Di)  了(Di)  小(VH13)
雨(Naa)  ，(COMMACATEGORY) .....
</P>
........
</TEXT>
</DOC>
```

Figure 2: An example with CKIP POS-tags

```
<DOC id="XIN_CMN_20010101.0004" type="story">
<HEADLINE>
印度/ns  平靜/ad  迎接/v  新/a  千年/t
</HEADLINE>
<DATELINE>
新华社/nt  新德里/ns  1 月/t  1 日/t  电/n
</DATELINE>
<TEXT>
<P>
(/w  记者/n  熊/nr  昌义/nr ）/w  10 亿/m  印度/ns  人/n  以/p  平
靜/a  的/u  心情/n  迎来/v  了/u  新/a  千年/t  。/w .....
 </P>
<P>
1 日/t   凌晨/t   ,/w   新德里/ns   虽/c   下/f   起/v   了/u   小雨
/n  ,/w ......
</P>
........
</TEXT>
</DOC>
```

Figure 3: An example with ICL POS-tags

4. Evaluation

4.1 Consistency of Segmentation

In order to measure the annotation quality of these two tagsets, evaluation criteria recall and precision were used to justify the documents of China's Xinhua News Agency from 2001 to 2004. The total number of documents is 303,493. MatchWord# means the number of words that two systems have agreement in terms of segmentation. And RefWord# is the number after segmentation. Table 1 shows the average evaluation results. The standard error between the documents is about o.o9.

Year	CKIP		ICL	
	Recall	Precision	Recall	Precision
2001	87.34	89.51	89.88	87.66
2002	84.96	86.62	87.26	84.95
2003	87.45	89.72	90.17	87.86
2004	87.25	89.63	90.06	87.64
All	86.84	88.98	89.45	87.13

Note:

$$\mathrm{Re}\,call(CKIP) = MatchWord\# \ / \ \ \mathrm{Re}\,fWord\#(ICL)$$

$$\mathrm{Pr}\,ecision(CKIP) = MatchWord\# \ / \ \ \mathrm{Re}\,fWord\#(CKIP)$$

$$\mathrm{Re}\,call(ICL) = MatchWord\# \ / \ \ \mathrm{Re}\,fWord\#(CKIP)$$

$$\mathrm{Pr}\,ecision(ICL) = MatchWord\# \ / \ \ \mathrm{Re}\,fWord\#(ICL)$$

Table 1 Evaluation for Segmentation

4.2 Corpus-based POS Tags Mapping

The POS-tagged documents of China's Xinhua News Agency from 2001 to 2004 were used for tags mapping. First, both versions of the tagged corpus were aligned in order to compare their segmentation results. This comparison shows that the two systems have about 85% agreement in terms of segmentation. Next, for all words where both system agrees in segmentation, we obtain mappings from the CKIP-AS tagset to the ICL-PKU tagset and ICL-PKU tagset to CKIP-AS tagset. There are 48 tags in CKIP system and 40 tags in ICL system. The main difference between these two tagsets is the CKIP POS tags is hierarchy design.

Table 2 and 3 show both all the possible mappings as well as their probabilities.It is easy to see from Table 2 that among the 48 CKIP pos tags, only 9 do not map to a clearly dominant ICL pos: Dfb, Dk, I, Nb, Nc, Ncd, Neqb T, and VH. All the other tags, to varying degrees, are mapped to one dominant corresponding pos tag with other less dominant mappings. It is interesting to note that Dfb, I, and T are minor categories without concrete semantic meaning, while Nc and Ncd are highly dependent on semantic interpretation. Among 40 ICL POS tags in Table 3, there are 15 do not map to one dominant corresponding pos. The high degree of correspondences, however, does confirm that the two linguistic systems are still very similar and that comparative studies based on these two different tagsets are valid.

CKIP Tag	PKU Mapping Tag
A (Non-predicative adjective)	b (53.6%)
Caa (Conjunctive conjunction)	c (84.8%)
Cab (Conjunction, e.g.等等)	u (87.5%)
Cba (Conjunction, e.g.的話)	u (92.4%)
Cbb (Correletive Conjunction)	c (82.9%)
D (Adverb)	d (67.5%)
Da (Quantitative Adverb)	d (88.5%)
DE (的,之,得,地)	u (91.5%)
Dfa (Pre-verbal Adverb of degree)	d (91.4%)
Dfb (Post-verbal Adverb of degree)	t (34.6%); q (24.9%)
Di (Aspectual Adverb)	u (93.3%)
Dk (Sentential Adverb)	v (49.1%); n (15.8%)
FW (Foreign Word)	m (82.5%)
I (Interjection)	e (34.7%); j (27.4%)
Na (Common Noun)	n (82.6%)
Nb (Proper Noun)	nr (47.6%); ns (10%)
Nc (Place Noun)	ns (47.2%); n (36.4%)
Ncd (Localizer)	f (45.4%); m (37.9%)
Nd (Time Noun)	t (88.3%)
Nep (Demonstrative Determinatives)	r (98.5%)
Neqa (Quantitative Determinatives)	m (55.1%)
Neqb (Post-quantitative Determinatives)	m (48%); a (32.7%)
Nes (Specific Determinatives)	r (55.7%)
Neu (Numeral Determinatives)	m (99.6%)
Nf (Measure)	q (90%)
Ng (Postposition)	f (76.7%)
Nh (Pronoun)	r (89.5%)
Nv (Verbal Nominalization)	vn (65.2%)
P (Preposition)	p (87.3%)
SHI (是)	v (95.2%)
T (Particle)	y (47.7%); u (43%)
VA (Active Intransitive Verb)	v (52.7%)
VAC (Active Causative Verb)	v (83.8%)
VB (Active Pseudo-transitive Verb)	v (59.3%)
VC (Active Transitive Verb)	v (70.8%)
VCL (Active Verb with a Locative Object)	v (82.7%)
VD (Ditransitive Verb)	v (73.2%)
VE (Active Verb with a Sentential Object)	v (85.1%)
VF (Active Verb with a Verbal Object)	v (82.9%)
VG (Classificatory Verb)	v (70.8%)
VH (Stative Intransitive Verb)	a (43.1%); v (16.3%)
VHC (Stative Causative Verb)	v (63.2%)
VI (Stative Pseudo-transitive Verb)	v (70.7%)
VJ (Stative Transitive Verb)	v (81.8%)
VK (Stative Verb with a Sentential Object)	v (88.4%)
VL (Stative Verb with a Verbal Object)	v (79.1%)
V_2 (有)	v (95%)
*CATEGORY (punctuation)	w (96.1%)

Table 2 CKIP to ICL Tag Mapping Table

4.3 Word Correction for Segmentation

We further analyzed the segmentation results between two systems. If a longer word is segmented by one system and another system divided into more than two words, the program automatically verified this kind of pattern and recorded the word pair and its frequency in the corpus. There are 88,443 word pairs in longer words with ICL system to shorter words with CKIP system, and 423,744 word pairs in longer words with CKIP system to shorter words with ICL system. Manual analysis of the frequency of word pairs is larger than 99 in ICL longer words indicated that about 82% of words, i.e. 2, 065 words of 2534 words, can be corrected in CKIP Segmentation. Among the remaining 469 words, 163 words belong to the differences between semantic meanings of words and segmentation criteria. For example, ICL system regarded "2008 年" as a word , however, in CKIP system, "2008" and "年" were two word that means a digit and an unit.

PKU Tag	ICL Mapping Tag
a (Adjective)	VH (75.5%)
ad (Adjective)	VH (72.6%)
an (Noun adjective)	VH (65.7%)
ag (Adjectival morpheme)	Caa (45.6%); VH (21%)
b (Distinguishing word)	DE (33.5%); A (29.2%)
c (Conjunction)	Caa (50.5%)
d (Adverb)	D (74.5%)
dg (Adverbial morpheme)	P (34.6%); D (12.6%); VJ (8 %)
e (Interjection)	DE (43.9%); T (15.7%)
f (Location)	Ng (58.3%)
g (Morpheme)	FW (52.1%)
h (Pre-adjective of degree)	A (21.8%); Nes (20%); Nc (14.9%)
i (Phrase)	VH (64%)
j (Abbreviation)	Nc (36.3%); Na (31.6%)
k (Post-adjective of degree)	Na (80.1%)
l (Idiom)	Na (35.9%); VH (28.4%)
m (Measure)	Neu (72.2%)
n (Noun)	Na (78.7%)
ng (Noun morpheme)	Na (34.9%); Ng (31.5%)
nr (Proper noun)	Nb (80.3%)
ns (Place noun)	Nc (92.2%)
nt (Affiliation)	Nc (74.4%)
nx (Non-Chinese character)	*CATEGORY (98.9%)
nz (Other special noun)	Nb (44.2%); Na (26.5%)
o (onomatopoeia)	D (36.4%); VC (17%)
p (Preposition)	P (86%)
q (Classifier)	Nf (90.2%)
r (Pronoun)	Nh (42.2%); Nep (27.8%)
s (Locational noun)	Nc (80.3%)
t (Time noun)	Nd (96.9%)
tg (Time morpheme)	Nd (54%)
u (Auxiliary)	DE (76.8%)
v (Verb)	VC (32.7%); VE (11.4%); VJ (8.2%)
vd (Adverbial verb)	VH (17.5%); VC (14%); VL (12.9%); D (10.8%)
vg (Verbal morpheme)	VC (22%); Na (19.3%); D (10.4%)
vn (Noun verb)	VC (34.7%); Na (28.2%)
w (Punctuation marks)	*CATEGORY (95.3%)
x (Non-morpheme)	FW (68.2%)
y (Modal particle)	T (61.3%)
z (Stative modifier)	VH (68.4%)

Table 3 ICL to CKIP Tag Mapping Table

5. Conclusion and Future Work

After the regular and default tag-to-tag mapping between CKIP and ICL systems are established based on the above data and manual analysis, we will investigate and exceptional mappings. Some of these mappings will be explained as non-homomorphism between the systems, yet others will be identified as potential tagging errors. We will investigate the possible error patterns when the segmented words were inconsistent. Once these error patterns were successfully found, models for automatic correction and estimation of confidence of automatic tags will be devised. An iteration algorithm that will improve the quality of both versions of the tagged corpus will be proposed and tested.

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*

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29 (3), pp. 333-347.

Ma, W.-Y. and Chen, K.-J. (2005). Design of CKIP Chinese Word Segmentation System. *Chinese and Oriental Languages Information Processing Society*, 14(3), pp. 235-249.

Ma, W.-Y. and Huang, C.-R. (2006). Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-5).*

Tseng, H. H. and Chen, K. J. (2002). Design of Chinese Morphological Analyzer. *Proceedings of 1st SIGHAN Workshop on Chinese Language Processing*.

Xia, F., Palmer, M., Xue, N., Okurowski, M. E., Kovarik, J., Chiou, F.-D., Huang, S., Kroch, T. and Marcus, M. (2000). Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2).*

Yu, S., Duan, H., Zhu, X. and Sun, B. (2002). The Basic Processing of Contemporary Chinese Corpus at Peking University- Specification. *Journal of Chinese Information Processing*, 16(5&6), pp. 49-64.

Yu, S., Duan, H., Zhu, X., Swen, B. and Chang, B. (2003). Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation. *Journal of Chinese Language and Computing*, 13(2), pp.121-158.

Zhang, H.-P., Liu, Q., Cheng, X.-Q., Zhang, Hao and Yu, H.-K. (2003a). Chinese Lexical Analysis Using Hierarchical Hidden Markov Model. *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing.*

Zhang, H.-P., Yu, H.-K., Xiong, D.-Y. and Liu, Q. (2003b). HHMM-based Chinese Lexical Analyzer ICTCLAS. *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing.*