Consistent Word Segmentation, Part-of-Speech Tagging and Dependency Labelling Annotation for Chinese Language

Mo Shen[1], Wingmui Li[2], Hyunjeong Choe[1], Chenhui Chu[3],
Daisuke Kawahara[4], Sadao Kurohashi[4]
[1] Google Inc., California, USA
[2] The Chinese University of Hong Kong, Shatin, Hong Kong
[3] Japan Science and Technology Agency
[4] Graduate School of Informatics, Kyoto University, Kyoto, Japan
{moshen|wjli|hyunjeongc}@google.com,
chu@pa.jst.jp, {dk|kuro}@i.kyoto-u.ac.jp

Abstract

In this paper, we propose a new annotation approach to Chinese word segmentation, part-of-speech (POS) tagging and dependency labelling that aims to overcome the two major issues in traditional morphology-based annotation: Inconsistency and data sparsity. We re-annotate the Penn Chinese Treebank 5.0 (CTB5) and demonstrate the advantages of this approach compared to the original CTB5 annotation through word segmentation, POS tagging and machine translation experiments.

1 Introduction

The definition of “word” is an open problem in Chinese linguistics. In previous studies of Chinese corpus annotation (Duan et al., 2003; Huang et al., 1997; Xia, 2000), the judgement of word-hood of a meaningful string is based on the analysis of morphology: A morpheme in Chinese is defined as the smallest combination of meaning and phonetic sound in Chinese language, which can be classified into two major types:

1). **Free morphemes**, which can either be words by themselves or form words with other morphemes; and

2). **Bound morphemes**, which can only form words by attaching to other morphemes.

An issue with word definition using morpheme classification is that, it potentially undermines the consistency of the representation of words. For example, “论” (theory) is a bound morpheme, therefore the string “进化论” (theory of evolution) is treated as a word; on the other hand the string “进化 | 理论” (theory of evolution) are treated as two words, despite the fact that the two strings have the same meaning and structure. In another example, “者” (person) is considered as a bound morpheme, therefore “反对自由贸易者” (people who are against free trade) is treated as one word, while the string without the bound morpheme, i.e. “反对 | 自由 | 贸易” (be against free trade), can only be treated as a phrase of three words.

The morphology-based word definition can also make the data sparsity problem worse in corpus annotation. As an evidence, in the Penn Chinese Treebank 5.0 (CTB5) which is an annotated corpus widely used to train Chinese morphological analysis systems, we found that one of the major sources of the out-of-vocabulary (OOV) words is the compounds that end with a monosyllabic bound morpheme. For example, compounds 利用率 (utility rate) and 次品率 (rate of defective product) end with the bound morpheme 率 (rate); 完成度 (degree of completion) and 活跃度 (degree of activity) end with the bound morpheme 度 (degree); 持续性 (sustainability) and 挥发性 (property of volatile) end with the bound morpheme 性 (property). While these compounds are sparse in the corpus, the morphemes which they

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 298–308, Osaka, Japan, December 11-17 2016.

POS Pattern	Example
pronoun + noun	我校 (this university)
locative + noun	后门 (back door)
locative + verb	前述 (described above)
noun + locative	室内 (indoor)
pronoun + locative	此外 (besides)
adverb + verb	猝死 (sudden death)
noun + noun	厂房 (factory plant)
noun + measure	车辆 (vehicles)
adjective + noun	佳酿 (wines)
adjective + measure	高层 (high level)
verb + verb	抽取 (extract)
verb + particle	写完 (finish writing)
verb + adjective	打碎 (break)
verb + locative	综上 (accordingly)
verb + noun	辞职 (resign)
adjective + adjective	优雅 (elegant)
adverb + adjective	最新 (latest)
determiner + noun	各界 (all walks of life)
determiner + temporal	翌日 (the next day)

Table 1. Disyllabic character-level POS patterns.

CTB5 Example	Re-annotation
副主席/NN (vice president)	副/JJ (vice) 主席/NN (president)
透明度/NN (transparency)	透明/JJ (transparent) 度/SFN (degree)
非生产性/NN (unproductiveness)	非/JJ (none) 生产/VV (produce) 性/SFN (property)
中央集权式/JJ (politically centralized)	中央/NN (center) 集权/NN (centralization) 式/SFA (type)

Table 2. Some examples of the word and POS annotation in the original CTB5 and our re-annotation.

consist of can be frequently observed; this means these OOV words can be observed and learnt by a word segmenter if we split the morphemes as individual words in the annotation.

In this paper, we propose a simple annotation approach for Chinese word segmentation that overcomes the two issues: inconsistency and data sparsity, which are found in the traditional morphology-based annotation approach. We further propose a tagset for part-of-speech tagging and a label set for dependency labelling, which are consistent with our word segmentation strategy and capture more Chinese-specific syntactic structures. We re-annotate the entire CTB5 using this approach, and through word segmentation, POS tagging and machine translation experiments we demonstrate the advantages of our annotation approach compared to the original approach adopted in CTB5.

The remainder of this paper is organized as follows: in section 2 we will describe our proposed annotation approach to word segmentation; in section 3 we will present a POS tagset which is consistent with our word segmentation strategy and a new dependency label set; in section 4 we will demonstrate the effectiveness of our approach compared to the original CTB5 through experiments; we will conclude our work in the last section.

2 Word Segmentation Annotation

We categorize the words in CTB5 into three categories: Common words, names, and idioms. For names and idioms, we keep them as individual words since their word boundaries are relatively easy to recognize and the consistency in manual annotation can be achieved with less efforts. We will mainly focus on describing the treatments of common words in this section.

Tag	Description	Count in CTB5	Proposed annotation
NN	Noun	134,321	137,816
PU	Punctuation	75,794	75,935
VV	Verb	68,789	75,033
AD	Adverb	36,122	35,922
NR	Proper Noun	29,804	30,985
P	Preposition	17,280	17,721
CD	Cardinal Number	16,030	21,493
M	Measure Word	13,668	18,091
JJ	Adjective	12,979	13,898
DEC	Complementizer	12,310	12,346
DEG	Genitive Marker	12,145	12,145
NT	Temporal Noun	9,467	4,524
LC	Locative	7,676	0
VA	Predicative Adjective	7,630	7,518
CC	Coordinating Conjunction	7,137	7,134
PN	Pronoun	6,536	6,646
DT	Determiner	5,901	5,970
VC	Copula	5,338	5,521
AS	Aspect Particle	4,027	4,033
VE	“you3” (“have”)	2,980	2,980
OD	Ordinal Number	1,661	1,661
MSP	Other Particles	1,316	1,316
ETC	“deng3” (“etc.”)	1,287	0
CS	Subordinating Conjunction	888	888
BA	Causative Auxiliary	751	756
DEV	Manner Marker	621	627
SP	Sentence-final Particle	466	466
SB	Short Passive Auxiliary	451	451
DER	Resultative Marker	258	258
LB	Long Passive Auxiliary	245	245
FW	Foreign Word	33	391
IJ	Interjection	12	17
X	Unknown	6	6
SFN*	Nominal Suffix	0	13,212
SFA*	Adjectival Suffix	0	438
SFV*	Verbal Suffix	0	129

Table 3. Proposed tagset for part-of-speech tagging. The underlined characters in the examples correspond to the tags on the left-most column. The CTB POS are also shown.

The key in our method to define the boundaries of common words is the character-level POS pattern. Character-level POS has been introduced in previous studies (Zhang et al., 2013; Shen et al., 2014) which captures the grammatical roles of Chinese characters inside words; we further develop this idea and use it as a criterion in word definition.

We treat a meaningful disyllabic strings as a word if it falls into one of the character-level POS patterns listed in Table 1. The reason we focus on disyllabic patterns instead of other polysyllabic ones is that, based on our observation, meaningful strings with 3 or more syllables (other than names and idioms) are always compounds in Chinese, and therefore can be segmented into a sequence of monosyllabic and disyllabic tokens based on their internal structures. On the other hand, the internal structure in a disyllabic token, though still exists, is more implicit and harder to describe with syntactical relations; we believe that it would increase the difficulties for subsequent tasks, such as dependency parsing, if we further segment these disyllabic strings.

Following this strategy, a polysyllabic word can be then segmented based on its structure. This is illustrated with examples in Table 2.

3 Part-of-Speech and Dependency Label Set

To perform POS tagging re-annotation on CTB5 together with our proposed word segmentation approach, we use a POS tagset which is derived from the one used in the original CTB5 annotation. We show the tagset in Table 3 with comparison of number of occurrences of each tag in the original CTB5 and the re-annotated version, respectively. The tagset introduces several changes: First, we eliminate the use of the "LC" tag for locative words. This tag is assigned to all words that indicate locations and directions, such as 上 (up), 下(down), 左 (left), 右 (right), 内 (inside), 外 (outside) etc.. We instead tag these words based on their real syntactic roles in sentences, such as "NN" (noun), "AD" (adverb) or "VV" (verb). Second, we add three new tags into the tagset for suffixes: "SFN" (nominal suffix), "SFA" (adjectival suffix), and "SFV" (verbal suffix). These tags are given to monosyllabic tokens appearing at the end of compounds, which are the bound morphemes in the traditional view. Based on our observation, these tokens have the ability to determine the syntactic role of the entire compound. For example, any compound that end with a nominal suffix "度" (degree) always act as nouns in a sentence. It should be noted that because of this characteristic of suffixes, we can tag the children of suffixes in compounds based on their meaning but not their syntactic roles. We show some examples in Table 2 to illustrate our POS tagging strategy for compounds.

In Table 4 we present a dependency label set developed based on the Stanford Dependencies (De Marneffe et al., 2006) and its Chinese version (Chang et al., 2009), which defines 45 dependency relations for Chinese sentences. This label set is also closely related to the Universal Dependency[1] with many of their labels compatible with each other. We explain the major characteristics of our label set in the following subsection.

3.1 Chinese Specific Labels

dislocated The label "dislocated" is originally defined in the *universal dependencies* for languages such as Japanese to describe the syntactic relation of words in a topic–comment structure, but is not defined for Chinese. However, in Chinese it is frequent to see the topic–comment structure in a sentence, for example:

1. 這/this 本/-measure- 書/book 他/he 買/buy 的/-particle- (This book, he bought it)

In this sentence, 这本书 (this book) is the topic and 他买的 (he bought) is the comment. One common view of the syntactic structure of this sentence is that, 他 (he) is the subject of the predicate 他 (buy), and 书 (book) is the direct object. This treatment sees a topic–comment structure as having an OSV (object-subject-verb) word order, which is acceptable; it however has some problems in certain cases, for example:

2. 這/this 本/-measure- 書/book 他/he 買/buy 的/-particle- 昨天/yesterday 不見/disappear 了/-particle- (This book that he bought disappeared yesterday)

In this sentence, 书 (book) is still the direct object of 买 (buy), while it is also the subject of 不见 (disappear). Because of the nature of the dependency grammar we adopted, for such a structure we would have to choose one relation for 书 (book), either "nsubj" or "dobj", and discard the other relation which would cause a loss of the syntactic information encoded in the parse tree.

Moreover, the OSV word order cannot explain all topic-comment structure such as the following example:

3. 這/this 場/-measure- 火/fire 幸虧/fortunately 消防/firefighting 隊/team 來/come 得/-particle- 早/early (This fire, fortunately the firefighters came in time)

[1] http://universaldependencies.org/

Label	Description	Example Phrase	Example Dependency
acomp	adjectival complement	鞋是全新的 (the shoes are brand new)	**acomp**(是 are, 全新 brand new)
advmod	adverbial modifier	他看上去很疲倦 (he looks very tired)	**advmod**(疲倦 tired, 很 very)
amod	adjectival modifier	漂亮的首飾 (cute accessory)	**amod**(首飾 accessory, 漂亮 cute)
appos	appositional modifier	總統奧巴馬 (president Obama)	**appos**(奧巴馬 Obama, 總統 president)
asp	aspect marker	他給了我一本書 (he gave me a book)	**asp**(給 gave, 了 -aspect-)
attr	attributive modifier	他是個醫生 (he is a doctor)	**attr**(是 is, 醫生 doctor)
aux	auxiliary verb	必須解決 (must solve)	**aux**(解決 solve, 必須 must)
auxpass	passive auxiliary	他被刺殺了 (he was assassinated)	**auxpass**(刺殺 assassinated, 被 -auxiliary-)
auxcaus	causative auxiliary	把問題解決(solve the problem)	**auxcaus**(解決 solve, 把 -auxiliary-)
cc	coordinating conjunction	聰明又可愛 (smart and cute)	**cc**(聰明 smart, 又 and)
ccomp	closed clausal complement	他說他喜歡游泳 (He said that he likes swimming)	**ccomp**(說 said, 喜歡 likes)
conj	conjunct	聰明又可愛 (smart and cute)	**conj**(聰明 smart, 可愛 cute)
csubj	clausal subject	能夠代表祖國參賽是他的夢想 (being able to play in the game for his country is his dream)	**csubj**(是 is, 參賽 play)
csubjpass	clausal passive subject	他在考試中作弊被老師發現了 (that he cheated during the exam is found out by the teacher)	**csubjpass**(發現 find out, 作弊 cheet)
dep	undefined dependency	添加一個日程安排時間星期二地點 3 樓 (add an event, time Tuesday, location 3rd floor)	**dep**(時間, 添加)
det	determiner	那本書 (that book)	**det**(本 -measure-, 那 that)
discourse	discoursal modifier	唉，終於到星期五了 (oh, thank God it’s Friday)	**discourse**(到 is, 唉 oh) **discourse**(到 is, 了-sentence-final particle-)
dislocated	dislocated modifier	書是他買的 (book he bought) 這場火幸虧消防隊來得早. (this fire, fortunately the firefighters came in time)	**dislocated**(書 book, 買 buy) **dislocated**(火 fire, 來 come)
dobj	direct object	買了一本書(bought a book)	**dobj**(買 bought, 書 book)
foreign	foreign compound	職棒大聯盟（Major League Baseball）	**foreign**(Major, League) **foreign**(Major, Baseball)
iobj	indirect object	他給了我一本書 (he gave me a book)	**iobj**(給 gave, 我 me)
list	list relation	添加一個日程安排時間星期二地點 3 樓 (add an event, time Tuesday, location 3rd floor)	**list**(時間 time, 地點 location)
mark	clause marker	他把信件給我之後就走了 (he left after he gave me the letter)	**mark**(給 give, 之後 after) **mark**(走 leave, 就 then)
mes	measure relation	一本書(a book)	**mes**(書 book, 本 -measure-)

ncomp	nominal complement	坐在椅子上 (sit on a chair)	**ncomp**(椅子 chair, 上 -complementizer-)
neg	negation modifier	不擅長 (be not good at)	**neg**(擅長 be good at, 不 not)
nn	noun compound modifier	原油期貨價格 (oil futures price)	**nn**(價格 price, 原油 oil) **nn**(價格 price, 期貨 futures)
npadvmod	noun phrase adverbial modifier	大約十米左右寬 (about 10 m wide)	**npadvmod**(寬 wide, 米 m)
nsubj	nominal subject	他給了我一本書 (he gave me a book)	**nsubj**(給 gave, 他 he)
nsubjpass	passive nominal subject	他被刺殺了 (he was assassinated)	**nsubjpass**(刺殺 assassinated, 他 he)
num	numeric modifier	一本書 (a book)	**num**(本 -measure-, 一 a)
p	punctuation	梨、橘子和香蕉 (pears, oranges, bananas)	**p**(梨 pears, 、)
pcomp	prepositional complement	由於路上人太多，我遲到了 (because it was so crowded, I was late)	**pcomp**(由於 because, 太多 so crowded)
pobj	prepositional object	他坐在椅子上 (he sits on a chair)	**pobj**(在 on, 椅子 chair)
ps	associative marker	這是我的家 (this is my home)	**ps**(我 me, 的 ‘s)
poss	possessive modifier	這是我的家 (this is my home)	**poss**(家 home, 我 me)
prep	prepositional modifier	他坐在椅子上 (he sits on a chair)	**prep**(坐 sits, 在 on)
prt	phrasal verb particle relation	他們打起來了 (they started a fight) 把數據整理成報告 (summarize the data into a report)	**prt**(打 fight, 起來 -auxiliary-) **prt**(整理 summarize, 成 become)
rcmodrel	relative clause complementizer	他回來的時候 (by the time he came back)	**rcmodrel**(回來 come back, 的 -complementizer-)
		這本是他買的書 (this is the book he bought)	**rcmodrel**(買 buy, 的 -complementizer-)
rcmod	relative clause modifier	他回來的時候 (by the time he came back)	**rcmod**(時候 time, 回來 come back)
		這本是他買的書 (this is the book he bought)	**rcmod(**書 book, 買 buy)
suff	suffix relation	科技界 (sci-tech industry)	**suff**(界 industry, 科技 sci-tech)
tmod	temporal modifier	他回來的時候天已經亮了 (it was bright outside by the time he came back)	**tmod**(亮 bright, 時候 time)
topic	topic marker	這本書是他買的 (this is the book he bought)	**topic**(書 book, 是 is)
		這是他們所不能想像的 (this is what they can’t imagine)	**topic**(這 this, 是 is)
vmod	verb modifier	他打開門發現屋裏有人 (he opened the door and found out there is somebody inside)	**vmod(**發現 found out, 打開 open)
xcomp	open clausal complement	他不喜歡打網球 (he doesn’t like to play tennis)	**xcomp**(喜歡 like, 打 play)

Table 4. Proposed dependency label set.

Unlike in the other two examples, the topic here, 這場火 (this fire), is not the direct object of the verb in the comment, 幸虧消防隊來得早 (fortunately the firefighters came in time).

To overcome these difficulties, we employ a different view which treats the topic–comment structure as having double subjects in a SSV word order. We define the first subject, 这本书 (this book) in example 2, as the head in a "dislocated" relation, and the subject-verb phrase, 他买的 (he bought) in example 2, as the modifier. The head in this dislocated relation can then form a "nsubj" (nominal subject) relation with the main predicate of the sentence, 不见 (disappear). Similarly, in example 3, the topic and the comment still form a dislocated relation even though the topic is not a direct object of the verb in the comment.

prt and ***prep*** We define the "prt" relation in two ways:

i. A relation between a verb and a particle. For example, 想像 (imagine) is the head in a "prt" relation of 所 (particle) in the sentence 這是他們所不能想像的 (this is what they can’t imagine).

ii. A relation between a verb and its succeeding complement. For example, 打掃 (clean) is the head in a "prt" relation of 完 (finish) in the sentence 房間打掃完了 (the room has been cleaned).

We use the "prt" relation in the second case to capture the predicate-complement structure in Chinese. The verb 完 (finish) in the second example above functions to complement the meaning of the main verb, 打掃 (clean), and the sentence is still grammatical when the complement verb is removed: 房間打掃了 (the room is cleaned).

The complement verb sometimes also functions as a coverb in a serial verb construction, which takes its own direct object. For example:

4. 把/-auxiliary- 數據/data 整理/summarize 成/become 報告/report (summarize the data into a report)

Here the two verbs 整理 (summarize) and 成(become) form a "prt" relation, while they are the heads of 數據 (data) and 報告 (report) in the "dobj" relation.

A difficulty with labeling "prt" is that, it can be easily confused with the "prep" (prepositional modifier) relation. For example, one can argue that 成 (become) is a preposition instead of a verb and should be tagged as IN, so that the relation between 整理 (summarize) and 成 (become) would be "prep". To overcome this ambiguity, we apply a simple test: If the phrase headed by the word with a VV vs. IN ambiguity can be moved to a position before the main verb, then this word is a preposition and a prepositional modifier of the main verb; otherwise it is a verb. Here since the phrase "成 報告" (into report) cannot be moved to the position before 整理 (summarize), it should in fact be a verb phrase, not a prepositional phrase.

suff We define the suffix relation in a compound which has a "stem-suffix" structure. The suffix word with a POS tag SFN, SFA, or SFV is the root of the subtree formed by the words in the compound. It has one and only one child in this subtree, which is the head of the "stem", and the dependency relation between them is labelled as "suff".

The motivation of employing the "suff" label is to relieve the data sparseness problem of word forms in annotated corpora. Compounds, especially those with a "stem-suffix" structure, is a major source of new words in Chinese language. These compounds, however, often share a set of suffix words which has a limited amount of instances. We think it is more effective for a parser to learn from features with word forms by treating the suffix words as the heads of compounds.

4 Evaluation

4.1 Re-annotated Corpus

We re-annotated the entire CTB5 with our proposed word segmentation and POS tagging annotation strategies. We further re-annotated 3,000 sentences which are randomly sampled from the training set of CTB5 using our proposed dependency label set. This re-annotated set is compared with the same sentences with the original annotation in a machines translation experiment in section 4.3.

	CTB5	Re-annotated
Number of tokens	493,938	516,581
Avg. token length	1.63	1.55
Ratio of unknown words	14.67%	12.82%
Ratio of unknown word-POS pairs	15.02%	13.28%

Table 5. Statistics of the original CTB5 and our re-annotated version.

(a) Word Segmentation Results

Corpus	P	R	F
Original	97.21	97.36	97.28
Re-annotated	97.97	97.56	97.76
Re-annotated-partial	97.68	97.63	97.65

(b) Joint Segmentation and POS Tagging Results

Corpus	P	R	F
Original	93.42	93.56	93.49
Re-annotated	94.55	94.16	94.35

Table 6. Experimental results for morphological analysis on CTB5.

To evaluate the consistency of our annotation, 4 trained annotators were divided into two equal groups to perform 2-way annotation on a small subset (first 100 sentences in files 301-325), and each pair of annotators were assigned with 50 sentences. The inter-annotator agreement is 99.10% for segmentation, 98.37% for POS tagging, and 95.62% for dependency labeling.

Table 5 shows some of the statistics of the original and the re-annotated CTB5. We split CTB5 in the same data division as in previous studies (Jiang et al., 2008a; Jiang et al., 2008b; Kruengkrai et al., 2009; Zhang and Clark, 2010; Sun, 2011). The training, development and test set have 18,089, 350 and 348 sentences, respectively. Compared to the original CTB5, the re-annotated training set has a lower percentage of unknown words and unknown word-POS pairs found in the corresponding test set. This is consistent with our observation that compounds with internal structures are one of the major sources of OOV words.

4.2 Morphological Analysis Experiments

We compared the performance of a state-of-the-art joint word segmentation and part-of-speech tagging system (Kruengkrai et al., 2009) on the original and our re-annotated CTB5. We used the position-of-character (POC) tagset and the baseline feature set described in (Shen et al., 2014).

We trained all models using the averaged perceptron (Collins, 2002), which is an efficient and stable online learning algorithm. The models applied on all test sets are those that result in the best performance on the dev sets. To learn the characteristics of unknown words, we built the system's lexicon using only the words in the training data that appear at least 2 times.

We use precision, recall and the F-score to measure the performance of the systems. Precision (P) is defined as the percentage of output tokens that are consistent with the gold standard test data, and recall (R) is the percentage of tokens in the gold standard test data that are recognized in the output. The balanced F-score (F) is defined as $\frac{2 \cdot P \cdot R}{P+R}$.

We compared the performance of the morphological analyzer on the original and the re-annotated CTB5. The results of the word segmentation experiment and the joint experiment of segmentation and POS tagging are shown in Table 6(a) and Table 6(b), respectively. Each row in these tables shows the performance of the same system trained on the corresponding corpus.

For "Re-annotated-partial" in Table 6(a), we applied a different setting in order to directly compare the annotation consistency and data sparsity between the two corpora: We used the training set from the re-annotated corpus to train the system but the test set from the original corpus in the evaluation. To make the evaluation meaningful, we added an extra criterion when calculating the precision and the

System	BLEU-4
Character	31.60
Original	31.46
Re-annotated	**32.08**

Table 7. Experimental results for Chinese-Japanese machine translation on ASPEC corpus using Moses system.

System	BLEU-4
Original	32.00
Re-annotated	**32.97**

Table 8. Experimental results for Chinese-Japanese machine translation on ASPEC corpus using KyotoEBMT system.

recall: If the outmost boundaries of a sequence (two or more) of output tokens are consistent with a token in the test set, we consider that the output correctly identifies this token in the test set.

The results show that, the morphological analyzer can obtain higher accuracies in both word segmentation (0.48 points absolute in F-score) and joint (0.86 points absolute in F-score) experiments. Furthermore, in the word segmentation experiment "Re-annotated-partial" where we mapped the output of the system which is trained using the re-annotated training data to the original CTB5 test set, the accuracy is significantly higher[2] than that of the "Original", which demonstrates the better consistency in our re-annotation corpus.

4.3 Machine Translation Experiments

To show that a morphological analysis system and a dependency parsing system can both benefit from our re-annotation, we conducted two sets of Chinese-to-Japanese machine translation experiments where a morphological analyzer and a dependency parser are used respectively.

The parallel corpus we used is the Chinese-Japanese part of the Asian Scientific Paper Excerpt Corpus (ASPEC)[3], containing 672k sentence pairs. We used 2,090 and 2,107 additional sentence pairs for tuning and testing, respectively.

In the first set of experiments, we segmented the Japanese sentences using JUMAN (Kurohashi et al., 1994), and the Chinese sentences using the same morphological analyzer described in the last subsection. For decoding, we used the state-of-the-art phrase based statistical machine translation toolkit Moses (Koehn et al., 2007) with default options. We trained the 5-gram language models on the target side of the parallel corpora using the SRILM toolkit[4] with interpolated Kneser-Ney discounting. Tuning was performed by minimum error rate training (MERT) (Och, 2003), and it was re-run for every experiment.

In the second set of experiments, we used the same morphological analyzers to segment and tag the POS of Japanese and Chinese sentences as in the first set. We further parsed the dependency structures of the Japanese sentences using KNP (Kawahara and Kurohashi, 2006), a lexicalized probabilistic dependency parser, and for the Chinese sentences we used a second-order graph-based parser proposed in (Shen et al., 2012). For decoding, we used the tree-to-tree example-based machine translation framework KyotoEBMT[5] (Richardson et al., 2015) with default options.

We report results on the test set using BLEU-4 score, which was evaluated using the multi-bleu.perl script in Moses based on Juman segmentations. The significance test was performed using the bootstrap resampling method proposed by Koehn (2004).

In Table 7 we compare the performance of three Moses models: In "Character" we used a simple segmentation strategy for the Chinese sentences where we treated each character as a token; in "Original" and "Re-annotated" we segmented the Chinese sentences using the corresponding models described in

[2] $p < 0.05$ in McNemar's test.

[3] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

[4] http://www.speech.sri.com/projects/srilm

[5] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KyotoEBMT

the last subsection. The results show, with the underlying machine translation system being the same, the segmenter trained with the original CTB5 failed to support the system to outperform the simple character-based segmentation, while on the other hand the system using the segmenter trained with our re-annotated CTB5 significantly outperformed both "Character"[6] and "Original"[7].

In Table 8 we show the result of the experiment with KyotoEBMT, a tree-to-tree machine translation system which requires unlabeled dependency annotation in the model training. 3,000 sentences with original and re-annotated dependency labels were used for training the parsers in "original" and "re-annotated" settings, respectively. The result shows that, the model "Re-annotated" which used the training set with the proposed annotation, it significantly outperformed[8] the baseline model "Original" by 0.97 point in BLEU-4 score.

5 Conclusion

We have proposed a new annotation approach for Chinese word segmentation, part-of-speech tagging, and dependency labelling. By re-annotating the CTB5 and conducting word segmentation, POS tagging and machine translation experiments, we have demonstrated that this approach has the advantages in achieving higher annotation consistency as well as less data sparsity, compared to the original annotation of CTB5. We couldn't show the comparison in dependency parsing experiments as we currently have only 3,000 annotated sentences; this experiment will be included in our future work.

Reference

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, pages 51-59.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In Proceedings of EMNLP, pages 1–8.

Daisuke Kawahara and Sadao Kurohashi. 2006. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. In Proceedings of the Human Language Technology Conference of the NAACL, pages 176–183.

HuiMing Duan, XiaoJing Bai, BaoBao Chan, and ShiWen Yu. 2003. Chinese word segmentation at Peking University. In Proceedings of the second SIGHAN workshop on Chinese language processing, pages 152-155.

ChuRen Huang, KehJiann Chen, FengYi Chen, and LiLi Chang. 1997. Segmentation Standard for Chinese Natural Language Processing. Computational Linguistics and Chinese Language Processing vol. 2, no. 2, August 1997, pages 47-62.

Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lü. 2008a. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-speech Tagging. In Proceedings of ACL.

Wenbin Jiang, Haitao Mi, and Qun Liu. 2008b. Word Lattice Reranking for Chinese Word Segmentation and Part-of-speech Tagging. In Proceedings of COLING.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of EMNLP 2004, pages 388-395.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion, Demo and Poster Session, pages 177-180.

[6] $p < 0.5$
[7] $p < 0.1$
[8] $p < 0.1$

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, YiouWang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An Error-Driven Word-Character Hybird Model for Joint Chinese Word Segmentation and POS Tagging. In Proceedings of ACL-IJCNLP, pages 513-521.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Nagao Makoto. 1994. Improvements of Japanese Morphological Analyzer JUMAN. In Proceedings of the International Workshop on Sharable Natural Language, pages 22-28.

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In Proceedings of LREC 2006, pages 449-454.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 160-167.

John Richardson, Raj Dabre, Chenhui Chu, Fabien Cromières, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. KyotoEBMT System Description for the 2nd Workshop on Asian Translation. In Proceedings of the 2nd Workshop on Asian Translation, pages 54-60.

Mo Shen, Daisuke Kawahara, and Sadao Kurohashi. 2012. A Reranking Approach for Dependency Parsing with Variable-sized Subtree Features. In Proceedings of 26th Pacific Asia Conference on Language Information and Computing, pages 308-317.

Mo Shen, Hongxiao Liu, Daisuke Kawahara, and Sadao Kurohashi. 2014. Chinese Morphological Analysis with Character-level POS Tagging. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 253–258.

Weiwei Sun. 2011. A Stacked Sub-word Model for Joint Chinese Word Segmentation and Part-of-speech Tagging. In Proceedings of ACL-HLT, pages 1385–1394.

Fie Xia. 2000. The Segmentation Guidelines for the Penn Chinese Treebank (3.0). http://www.cis.upenn.edu/~chinese/segguide.3rd.ch.pdf.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese Parsing Exploiting Characters. In Proceedings of ACL, page 125-134.

Yue Zhang and Stephen Clark. 2010. A Fast Decoder for Joint Word Segmentation and POS-tagging Using a Single Discriminative Model. In Proceedings of EMNLP, pages 843–852.