A Dependency Treebank of the Chinese Buddhist Canon

Tak-sum Wong, John Lee
Halliday Centre for Intelligent Applications of Language Studies
Department of Linguistics and Translation, City University of Hong Kong
E-mail: tswong-c@my.cityu.edu.hk, jsylee@cityu.edu.hk

Abstract

We present a dependency treebank of the Chinese Buddhist Canon, which contains 1,514 texts with about 50 million Chinese characters. The treebank was created by an automatic parser trained on a smaller treebank, containing four manually annotated sutras (Lee and Kong, 2014). We report results on word segmentation, part-of-speech tagging and dependency parsing, and discuss challenges posed by the processing of medieval Chinese. In a case study, we exploit the treebank to examine verbs frequently associated with Buddha, and to analyze usage patterns of quotative verbs in direct speech. Our results suggest that certain quotative verbs imply status differences between the speaker and the listener.

Keywords: dependency treebank; Chinese Buddhist canon; quotative verbs

1. Introduction

Over the past decade, there has been growing interest in building treebanks for historical texts, not only for facilitating their reading but also for studying the historical languages in which they were written. The sacred texts of many major religious, for example, have been syntactically analysed: treebanks are now available for the Hebrew Bible (Wu & Lowery, 2006), the New Testament in Greek (Haug & Jøhndal, 2008), and the Qur'an in Classical Arabic (Dukes & Buckwalter, 2010).

With about 50 million characters, the sheer volume of the Chinese Buddhist Canon makes it difficult for any individual to perform manual analysis over the entire corpus. Although digitized versions of the Canon have enabled n-gram and other lexical analyses (Lancaster, 2010), it remains difficult to examine patterns in part-of-speech (POS) and sentence structures without syntactic annotations.

To date, the only treebank with Buddhist Chinese material consists only of four sutras (Lee and Kong, 2014). We trained a dependency parser on this small treebank, and then automatically parsed the entire Chinese Buddhist Canon. In this paper, we start with an overview of existing treebanks for ancient Chinese (Section 2). We then report the procedure for constructing this treebank, and evaluate its accuracy (Section 3). Finally, as a case study, we examine the verbs used by Buddha and other characters in the treebank, focusing on quotative verbs (Section 4).

2. Previous Work

Among large, diachronic corpora in ancient Chinese are the Academia Sinica Ancient Chinese Corpus (Wei et al., 1997) and the Sheffield Corpus of Chinese (Hu et al., 2005), both covering a wide range of time and genres. Although word-segmented and POS-tagged, they have not been syntactically analyzed.

To the best of our knowledge, only three treebanks are available to-date for ancient Chinese. First, a constituent-based treebank has been constructed on 1,000 sentences from pre-Qin texts (Huang et al., 2002). Second, a dependency treebank has been annotated for 32,000 characters of Tang poems, selected from the works of three poets from the 8th century CE (Lee and Kong, 2012). Third, and most related to this work, is a dependency treebank of four sutras, numbering about 50,000 characters, taken from the Chinese Buddhist Canon (Lee and Kong, 2014). Written in medieval Chinese, the Canon consists of translations of Buddhist texts from Indic languages, produced from the 2nd to the 11th centuries CE.

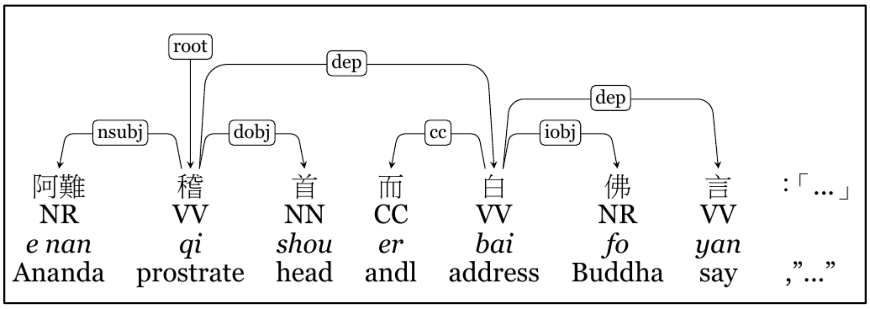

Figure 1: Dependency tree for the sentence, "Ananda bowed and addressed Buddha, saying, '…'".

3. Treebank Construction

3.1 Textual material

Our treebank is based on a digital version of the *Tripiṭaka Koreana*, the Korean Edition of the Chinese Buddhist Canon (Lancaster, 2010). This edition is derived from the most complete set of available printing blocks, those currently stored at Haein Monastery in Korea (Lancaster and Park, 1979).

3.2 Punctuation

The *Tripiṭaka Koreana* has no punctuation. In order to facilitate automatic syntactic analysis, we inserted punctuation from another digital edition of the Chinese Buddhist Canon, the *Taishō Revised Edition*, provided by the Chinese Buddhist Electronic Text Association (CBETA).

Although this version was derived from the same set of printing blocks as the *Tripiṭaka Koreana*, it does not represent the whole of the text glyphs found in the blocks. When the *Taishō Revised Edition* was produced in the 19th century, only 10,000 characters were available to the publishers and thus many substitutions of similar characters had to be made. In contrast, the digital version of the *Tripiṭaka Koreana* reproduced every glyph found in the blocks, making it more accurate for our purposes.

3.3 Training Data

Since off-the-shelf parsers are mostly intended for modern Chinese, we used a small dependency treebank of Buddhist Chinese texts (Lee and Kong, 2014) to train a word segmenter, POS tagger and dependency parser. We now describe the word segmentation method, POS tagset, and dependency relations adopted by our training treebank.

For word segmentation, the treebank largely adopted the guidelines for the Penn Chinese Treebank (Xue et al., 2005). The treebank also adopted the POS tagset of the Penn Chinese Treebank; however, since that tagset was originally developed for Modern Chinese, Lee and Kong had introduced some minor adaptations (2014).

As for dependency relations, the treebank followed Lee and Kong (2012) in adapting the Stanford Dependencies for Modern Chinese (Chang et al., 2009). It added five new relations, and made minor changes in the definitions of a number of relations.

3.4 Automatic Parsing

3.4.1. Word Segmentation

As the first step, we built a word segmenter with CRF^{++} (Lafferty, 2001) using the approach proposed by Zhao et al. (2007), which exploits the unigram, bigram, jump, punctuation and digital features, as well as external dictionaries. Compared to modern Chinese, fewer words in medieval Chinese text contain more than two syllables. Therefore, we followed Peng et al. (2004) and Tseng et al. (2005) in adopting a 2-tag set for word segmentation. Three Buddhist lexicons — the Soothill-Hodous Dictionary of Chinese Buddhist Terms (Soothill-Hodous & Lewis, 1995), the Person and Place Authority Databases from Dharma Drum Buddhist College (DDBC, 2008a; 2008b) — served as dictionaries.

Method	Precision	Recall	F Measure
CRF with external dictionaries	96.90%	98.28%	97.58%
CRF without external dictionaries	95.17%	97.96%	96.54%
Forward Maximal Matching	97.27%	95.83%	96.54%

Table 1: Word segmentation results on 10-fold cross validation of the treebank (Lee & Kong, 2014)

Table 1 shows the word segmentation results. The CRF model, in conjunction with the external dictionaries, yielded the best results. Some errors were due to the lack of coverage of the dictionaries for non-religious words; others resulted from ambiguity between word and phrase. For instance, 滅度 *mièdù* can be interpreted as one word meaning "nirvana" in some contexts, but also as a sequence of two coordinated verbs 滅 *miè* and 度 *dù* meaning "to extinguish and to save" in others. For another frequent expression, 如是 *rúshì*, the two characters as a whole serve as an adverb meaning "thus; so it is". When considered as two words, however, they form the phrase "like this" from 如 *rú* 'like' and 是 *shì* 'this'.

The dictionaries sometimes disagree on whether to include a common noun as part of the proper noun. For example, 舍衞國 *lit.* 'Śrāvastī country' is an entry in DDBC (2008b) but 舍衞 'Śrāvastī' *per se* is an entry in Soothill-Hodous and Lewis (1995). Segmentation results were also affected by limited coverage of the dictionaries for non-religious words.

3.4.2. Part-of-speech tagging

Next, we trained a part-of-speech (POS) tagger with CRF^{++} (Lafferty et al., 2001). In addition to the standard unigram and bigram features, we also used two external dictionaries — the DDBC Person and Place Authority Databases (DDBC, 2008a; 2008b), and a list of Sanskrit-transliterated terms harvested from Chu (1996, 1998, 1999) — to help the tagger recognize unseen nouns.

Method	Accuracy
CRF without external dictionaries	81.81%
CRF with DDBC	81.85%
CRF with DDBC + Chu	81.86%

Table 2: Results for POS tagging on 10-fold cross validation on the treebank (Lee & Kong, 2014)

Table 2 tabulates POS tagging results. The use of external dictionaries only slightly improved performance. The boundary between verbs and nouns in medieval Chinese, a inflection-poor language, are often not clear-cut; for example, the word 誦 *sòng* can serve as the verb 'to murmur' or the noun 'chant'. To complicate matters, the long history of the composition of the Canon — translations from the original Indic languages spanning over a millennium — means that the Chinese language itself varied within the Canon.

A few common words with multiple POS are responsible for many errors. For instance, the word 是 *shì* originally served only as the pronoun 'this', but later also took on the role of the copula 'to be' (Wang, 1998). Consider the sentence 是心是佛 *shì xìn shì fó* 'This mind is Buddha'. The same word *shì* occurs two times. In its first occurrence, it serves as the determiner for *xìn* 'mind', and should be tagged as "DT". In its second occurrence, however, it serves as the copula 'is', and should be tagged as VC.

Similarly, the words 若 *ruò*, and 如 *rú*, originally a verb 'be like', began to play the role of the conjunction 'if' and the preposition 'such as', respectively, in more recent texts. Finally, the word 者 *zhě* also evolved from a noun 'person' to also serve as a sentence-final particle. When a sentence ends with a noun containing *zhě* 'person' (e.g., *wén fǎ zhě* 聞法者 'anyone who hears the Dharma'), it is often difficult to distinguish between the two usages.

3.4.3. Dependency parsing

Lastly, we trained a Minimum-Spanning Tree parser (McDonald et al., 2006) to automatically infer dependency structures. Very few sentences in our training data are non-projective; we used the Eisner algorithm for projective parsing.

Word segmentation	POS-tagging	UAS	LAS
gold standard	gold standard	79.36%	74.60%
CRF	gold standard	66.16%	55.39%
CRF	CRF	61.24%	51.42%

Table 3: Unlabeled and labeled attachment scores for dependency parsing on 10-fold cross validation of the treebank (Lee & Kong, 2014)

Table 3 lists the parsing results. Parsing is challenging for medieval Chinese, an analytic rather than inflectional language. There are four main sources of error. First, in a serial verb construction, there is often confusion between the dependent (`dep`) and clause complement (`ccomp`) relations. Second, an indirect object (`iobj`) is often mistaken as a direct object (`dobj`), since the latter appears much more frequently. Third, the external object (`exd`), which marks vocatives, is often parsed as noun subject (`nsubj`). When a personal name occupies the sentence-initial position, the choice between these two options often depends on semantics. Fourth, when a verb appears before a noun, it is sometimes difficult to tell whether the noun modifies the verb as a direct object (`dobj`), or the verb modifies the noun (`vmod`).

4. Case Study

The syntactic information provided by the treebank can potentially help investigate a wide range of linguistic research topics on the Chinese Buddhist Canon. As a preliminary case study, we start by analyzing the verbs that are associated with Buddha.

Frequent verbs for Buddha. Table 4 lists the verbs for which 佛 *fó* 'Buddha', or one of his ten epithets, most frequently serve as the noun subject. Three of the top five verbs — *yán* , *shuō*, and *gào… yán* — are saying verbs. Their dominance reflects the sutras as the remembered words of Sakyamuni Buddha. The locations where Buddha delivered his sutras are usually recorded, hence the frequency of the verb 在 *zài* 'at'. He is often said to be "unhindered" or 無礙 *wú ài*, literally 'have no obstacle', contributing to the frequency of *wú*.

Verb	English	Percentage
言 *yán*	'to say'	9.6%
說 *shuō*	'to say'	6.0%
告……言 *gào… yán*	'to tell'	5.8%
無 *wú*	'to have not'	1.9%
在 *zài*	'be at; dwell in'	1.4%

Table 4: Five most frequent verbs for which Buddha is the subject

Quotative verb	English
言 *yán*	'to say'
告 *gào*	'to tell; to announce to'
白……言 *bái… yán*	'to address … and say'
說 *shuō*	'to say'
答曰 *dáyuē*	'to reply and say'
曰 *yuē*	'to say'
問 *wèn*	'to inquire'
答言 *dáyán*	'to reply and say'
告……言 *gào… yán*	'to tell… and say'
作 *zuò*	'to make'

Table 5: Ten most frequent quotative verbs

Quotative verbs. Probing further into the saying verbs, we now investigate their role in reporting direct speech. More specifically, we wish to investigate which ones are used as quotative verbs, i.e., verbs that introduce quoted speech (cf., 'said', 'tell' in English); and whether there are selectional differences among the characters according to their status.

In order to retrieve direct speech from texts, we identified all sentences enclosed within quotation marks. We then searched for the quotative verb associated with the quoted speech. Typically, the subject of this verb, or a coordinated verb, is the speaker; and the indirect object is the listener. For example, in Figure 1, 白 *bái* 'to address' is the quotative verb, Buddha is the listener, and his disciple Ānanda is the speaker. The most frequent quotative verbs are shown in Table 5.

If a verb indicates relative status between the speaker and listener, it should be used predominantly in one direction only; i.e., only when character X spoke to Y, but not when

Y spoke to X. In order to test this hypothesis, we retrieved all pairs of characters who spoke to one another at least 5 times. We then examined if any of the verbs in Table 5 tended to be used only in one direction within these pairs.

Two verbs stood out. In 95.5% of the pairs, only one character used 白 *bái* 'to address' to talk to the other, but not in the reverse direction. In 87.3% of the pairs, a similar trend held for 告 *gào* 'to tell'. These figures suggest that the choice of *bái* and *gào* is strongly influenced by the identities of the speaker and listener.

The statistics for Buddha further clarify the status difference implied by these verbs. When Buddha spoke to another person, he never used *bái*; this confirms that *bái* is "used by an inferior to address a superior" (Kieschnick, 2015:95). Conversely, when he listened, the speaker never addressed him with *gào*. It appears, then, that *gào* is reserved for speaking to someone of lower status.

5. Conclusions and Future Work

We have presented a dependency treebank of the Chinese Buddhist Canon based on the Korean Edition, the *Tripiṭaka Koreana*. The treebank was created by an automatic parser trained on a smaller treebank, containing four manually annotated sutras (Lee and Kong, 2014). We have reported results on word segmentation, POS tagging and dependency parsing, and discussed challenges posed by the processing of medieval Chinese.

In a case study, we have exploited syntactic information in the treebank to examine verb usage in the Canon. Focusing on verbs that report direct speech, our results confirmed that the verb *bái* is used by an inferior to address a superior, and found that the verb *gào* is used in the opposite direction, in a highly predictable manner.

For future work, we intend to improve the accuracy of POS tagging and dependency parsing, for example with active learning. We plan to exploit the treebank, on the one hand, as a research tool to perform "distant reading" (Moretti, 2013); and on the other hand, as a pedagogical tool to help students in "close reading". The medieval Chinese in the Canon often poses difficulty to speakers of modern Chinese. The POS and syntactic annotations can be expected to help readers digest the texts better and faster.

6. Acknowledgements

This work was partially supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. CityU 155412).

7. Bibliographical References

Chang, P.-C., Tseng, H., Jurafsky, D. and Manning C. D. (2009). Discriminative reordering with Chinese grammatical relations features. In *Proc. 3rd Workshop on Syntax and Structure in Statistical Translation*.

Chu, C.-n. (1996). *Vocabulary of Buddhist Sutras of the West Chin Dynasty*. Technical Report, National Chung Cheng University, Taipei.

Chu, C.-n. (1998). *The Lexicology of Buddhist Sutra in Ancient China (II) — Three Kingdoms*. Technical Report, National Chung Cheng University, Taipei.

Chu, C.-n. (1999). *The Lexicology of Buddhist Sutra in Ancient China (III) — Eastern Han Dynasty*. Technical Report, National Chung Cheng University, Taipei.

DDBC. (2008a). Buddhist Studies Person Authority Databases (Beta Version). Buddhist Studies Authority Database Project, Dharma Drum Buddhist College.

DDBC. (2008b). Buddhist Studies Place Authority Databases (Beta Version). Buddhist Studies Authority Database Project, Dharma Drum Buddhist College.

Dukes, K., Buckwalter T. (2010). A dependency treebank of the *Quran* using traditional Arabic grammar. In *Proceedings of the 7th International Conference on Informatics and Systems (INFOS)*, Cairo, Egypt, March 2010.

Haug, D., Jøhndal, M. (2008). Creating a parallel treebank of the old Indo-European Bible translations. In K. Ribarov, & C. Sporleder (Eds.), *Proceedings of the LREC Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, Marrakech, Morocco, June 2008.

Hu, X., Williamson, N. and McLaughlin, J. (2005). Sheffield Corpus of Chinese for diachronic linguistic study. *Literary and Linguistic Computing*, 20(3): 281-93.

Huang, L., Peng, Y., Wang, H. and Wu, Z. (2002). PCFG Parsing for Restricted Classical Chinese Texts. In *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, Taipei, Taiwan.

Kieschnick, J. (2015). A Primer in Chinese Buddhist Writings (vol. 1). http://religiousstudies.stanford.edu/a-primer-in-chinese-buddhist-writings/

Lafferty, J., McCallum A. and Pereira F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In *Proc. of ICML*, pp.282-289.

Lancaster, L. (2010). From Text to Image to Analysis: Visualization of Chinese Buddhist Canon. *Proceedings of Digital Humanities*. London, UK.

Lancaster, L., Park, S. (1979). *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley: Berkeley University Press.

Lee, J., Kong, Y. H. (2012). A dependency treebank of Classical Chinese poems. In R. Mihalcea, J. Chai, & A. Sarkar (Eds.), *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (HLT-NAACL)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 191-199.

Lee, J., Kong, Y. H. (2014). A dependency treebank of Chinese Buddhist texts. *Digital Scholarship in the Humanities* (doi: 10.1093/llc/fqu048).

McDonald, R., Lerman, K. and Pereira F. (2006). Multilingual dependency parsing with a two-stage discriminative parser. In *Proc. 10th Conference on Computational Natural Language Learning (CoNLL-X)*.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Peng, F., Feng, F. and McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields. In *Proceedings of COLING 2004*, Geneva, Switzerland, August 2004.

Soothill-Hodous, W. E., Lewis, H. (1995). *A Dictionary of Chinese Buddhist Terms: With Sanskrit and English Equivalents and a Sanskrit-Pali Index*. Surrey: Curzon Press.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D. and Manning, C. (2005). A conditional random field word segmenter for SIGHAN Bakeoff 2005. In C.-R. Huang, & G.-A. Levow (Eds.), *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea, October 2005.

Wang, L. (1998). *Classical Chinese* 古代漢語 [in Chinese]. Beijing: Zhonghua Press.

Wei, P.-C., Thompson, P. M., Liu, C.-H., Huang, C.-R., and Sun, C. (1997). Historical corpora for synchronic and diachronic linguistics studies. *Computational Linguistics and Chinese Language Processing*, 2(1):131-45.

Wu, A., Lowery, K. (2006). A Hebrew tree bank based on cantillation Marks. In *Proc. LREC*.

Xue, N., Xia, F., Chiou, F.-D. and Palmer, M. (2005). The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11, pp. 207-238.

Zhao, H., Huang, C.-N. and Li, M. (2007). An Improved Chinese Word Segmentation System with Conditional Random Field. In H. T. Ng, & O. O. Y. Kwong (Eds.), *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, pp. 162-165.