Treebanking for Data-driven Research in the Classroom

John Lee, Ying Cheuk Hui, Yin Hei Kong
Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
{jsylee,yingchui,yhkong}@cityu.edu.hk

Abstract

Data-driven research in linguistics typically involves the processes of data annotation, data visualization and identification of relevant patterns. We describe our experience in incorporating these processes at an undergraduate course on language information technology. Students collectively annotated the syntactic structures of a set of Classical Chinese poems; the resulting treebank was put on a platform for corpus search and visualization; finally, using this platform, students investigated research questions about the text of the treebank.

1 Introduction

Treebanks are now increasingly used as pedagogical tools (Crane et al., 2012), chiefly in two ways. On the one hand, in linguistics courses, students may use existing treebanks to perform quantitative analysis on syntactic patterns. On the other, in language courses, students may annotate syntactic structures to reinforce grammatical concepts, creating new treebanks. In this paper, we describe our experience in integrating these two processes into a research project in an undergraduate course, and discuss its benefits and challenges.

The project formed part of a course entitled "Language Information Technology". With no previous training, students collectively annotated the dependency structures of a portion of the *Three Hundred Tang Poems*, a popular anthology of Classical Chinese poems. The instructor edited the annotations, compiled them into a dependency treebank, and made it available for search and visualization on a web-based interface. Then, in a research assignment, students tackled questions on Chinese poetry with this treebank, which they had created with their own hands.

Combining the creation of a treebank with its use in a research assignment has many benefits. With respect to pedagogy, the assignment demonstrates to students the practical rationale for treebanks; the treebanking exercise familiarized students with the data and annotation scheme, helping them perform better on the assignment. With respect to longer-term effects, students perceive their own, tangible contribution to a field of scholarly research, in the form of linguistic annotations that are reusable by other scholars. The hands-on practice of a novel research methodology --- data-driven study in linguistics and literature --- should encourage them to apply it in their future fields of study.

The rest of the paper is organized as follows. Section 2 outlines previous use of treebanks in the classroom. Section 3 describes how our course was structured. Section 4 explains how students created the treebank, which formed the basis of the research assignment discussed in section 5. Section 6 presents the lessons learned and concludes.

2 Previous Work

Many current systems support the use of linguistic corpora for teaching and learning. One of many examples, the Visual Interactive Syntax Learning (VISL) system allows students to search, view, construct and label parse trees (Bick, 2005). The GATE system similarly facilitates corpus annotation, but it can also perform a variety of NLP tasks including POS tagging and parsing (Bontcheva et al., 2002).

These systems facilitate pedagogical use of treebanks in two main ways. First, students visualize parse trees and search for linguistic structures on existing treebanks. These functions

Proceedings of the Fourth Workshop on Teaching Natural Language Processing, pages 56–60,
Sofia, Bulgaria, August 4-9 2013.

support empirical and quantitative analysis of linguistic phenomena. Second, students also use their editing environment to create new dependency annotations on text, as exercises in learning a new language. The resulting treebank can then be made available for all scholars.

The latter type of usage has been implemented in Classics courses at six American universities. Students made dependency annotations on a Latin or Greek text, which the instructor then reconciled. The results contributed to the Latin and Ancient Greek Dependency Treebanks that are being compiled at the Perseus Project. In a study on 13 students, who had received limited training, the inter-annotator accuracy averaged 54.5% (Bamman & Crane, 2010).

Treebanking itself has also been taught in a course (Volk et al., 2005). Another notable case where students collectively created new linguistic resources has been reported at a graduate course in multilingual grammar engineering (Bender, 2007). Each student developed a grammar for automatic parsing of a new language. Over time, students' work was found to be effective in bringing feedback to the core grammar, and to facilitate empirical research on cross-linguistic comparisons.

A significant novelty in our course design is that, after students create new annotations for a treebank, they share the data with the rest of the class, and apply the freshly compiled treebank for linguistic research. We now describe how these two processes were implemented.

3 Course Structure

The project described in this paper was integrated into "Language Information Technology", an undergraduate course offered at the Department of Chinese, Translation and Linguistics at City University of Hong Kong. In the past semester, 44 students were enrolled. All majored in the Chinese language. As can be expected in a humanities department, the students had no technical background or experience in natural language processing. While some had previously taken linguistics courses, none was familiar with dependency grammar or its annotation scheme.

The course lasted for 13 weeks; weekly meetings consisted of a one-hour lecture and a two-hour tutorial or practicum. Roughly one half of this course was devoted to the treebanking project. In the first week, part-of-speech (POS) tagging was introduced, with English as the example language. During the practicum, students reviewed POS concepts with exercises and Stanford's online tagger[1]. In the second, dependency trees were introduced, again using examples in English. Lectures in the third and fourth weeks turned the attention to Chinese POS and dependency trees, using respectively the schemes defined at the Penn Chinese Treebank (Xue et al., 2005) and Stanford (Chang et al., 2009). During the practicums, adaptations to these schemes for Classical Chinese (Lee, 2012; Lee & Kong, 2012) were presented. In the fifth week, the web interface for searching and visualizing treebanks, which would later be used for a research assignment (see section 5), was demonstrated. Also, students were assigned individual texts for POS tagging and dependency labeling (see section 4). The practicum was devoted to discussions on difficulties in annotation.

The annotations were due two weeks later. After editing by the instructor, the treebank was posted on the aforementioned web interface, and the assignment was released. Overall, each student received 15 hours of class time in preparation for the treebanking project.

4 Treebank Annotation

The first task of the students, described in this section, is to annotate dependency structures of a set of Classical Chinese texts. The newly created treebank would then be used in a second task, to be discussed in the next section.

4.1 Choice of Material

Among the various literary genres, poetry enjoys perhaps the most elevated status in the Classical Chinese tradition. 320 poems from the Tang Dynasty, considered the golden age for poetry, have been grouped together in an anthology referred to as the *Three Hundred Tang Poems*. This anthology is perhaps the most well-known in the canon of Classical Chinese literature, and is featured without exception in the Chinese curriculum in secondary schools.

For the treebanking project, this corpus is ideal because it is both non-toy and not prohibitively difficult. As well-known literary heritage, this corpus lends interesting and significant questions to the research assignment (section 5). Moreover, unlike many other Chinese Classics, these poems are relatively simple to analyze, with each line containing not more than 7 characters. All students can be expected to have previous expo-

[1] http://nlp.stanford.edu:8080/parser/

sure to some of the poems. Finally, since the text is of such central importance, the resulting treebank is likely to be relevant to other scholars. It is especially motivating for students that their efforts would have an impact long after they receive their grades for the course.

4.2 Annotation Set-up and Results

Each of the 44 students was assigned four different poems from the *Three Hundred Tang Poems* for annotation, with a total of 144 characters.

The instructor manually corrected the student annotations. Using the corrected version as gold standard, the students achieved 68.1% labeled attachment score (LAS)[2]. The quality of individual students' annotations varied widely, from the lowest LAS at less than 10%, to the top student who scored more than 95%. Students were allowed to discuss their annotations with the instructor, but the correct annotations were never disclosed to them.

Part-of-speech tagging. The students achieved 93.9% accuracy for POS tagging, which compares reasonably with the agreement rate of 95.1% among two annotators reported on similar texts in (Lee, 2012). The tags with the highest error rates are shown in Table 1. The most frequent pairs of confusion are among the tags VA (predicative adjectives), AD (adverbs), and JJ (attributive adjectives).

The lack of morphology in Classical Chinese likely contributed to the confusion between AD and JJ. Consider the phrase 閒/AD *xian* 'relaxed' 坐/VV *zuo* 'sit', the first two characters from the line 閒坐說玄宗 "while sitting relaxedly, [we] gossip about Emperor Xuan". Here, the word *xian* 'relaxed' is an adverb describing the manner of *zuo* 'sit'; however, the same form can also serves as an adjective, perhaps leading a student to tag it as JJ.

Even more frequent is the confusion between JJ and VA. A typical example is the phrase 燭/NN *zhu* 'candle' 影/NN *ying* 'shadow' 深/VA *shen* 'becomes dark', the last three characters in the line 雲母屏風燭影深 "the shadow of the candle on the mica screen becomes dark". Despite hints from the word order, the student mistakenly considered *shen* 'becomes dark' as an attributive, rather than predicative, adjective.

[2] As a comparison, two heavily trained annotators achieved 91.2% agreement on similar texts (Lee and Kong, 2012), and performance of automatic parsers can reach LAS at 75.6% (Lee and Wong, 2012).

Tag	Error rate	Tag	Error rate
AD	20.1%	M	13.8%
P	20.0%	LC	9.4%
VA	19.1%	CD	6.6%
VC	16.1%	JJ	4.4%
PN	11.9%		

Table 1. POS tags with the highest error rates.

Head selection. Among those characters whose POS tags are correctly labeled, head selection is correct 81.8% of the time. As shown in Table 2, among the various POS, students most frequently had difficulty selecting heads for verbs. While there was a wide range of different kinds of mistakes, the most common one is to mistakenly take a noun as head, using the dependency label vmod (verb modifier).

Series of adverbs (AD) also turned out to be problematic; a third of the errors with AD fell into this case. Consider the two adverbs *bu* and *fu* in the phrase 不/AD *bu* 'not' 復/AD *fu* 'again' 返/VV *fan* 'return', the last three characters in the line 黃鶴一去不返 "once the crane leaves, it will not return". By convention in the Stanford framework (Chang et al., 2009), the head of the first adverb, *bu*, is the verb *fan* and not its adverb neighbor to the right, *fu*. This sort of error may be considered technical mistakes, rather than genuine misunderstanding of syntactic structure.

Tag	Error rate	Tag	Error rate
VV	28.9%	PN	9.6%
AD	10.0%	CD	7.1%
NR	9.8%	JJ	4.6%

Table 2. POS tags with the highest head selection error rates. The top three tags, CC, AS and SP, were omitted due to small sample size (only 3 each).

Dependency labels. When a wrong head is selected, the label was almost always also wrong. Among those words with the correct head, the accuracy in dependency labeling was 88.6%. Table 3 lists the labels with the lowest accuracy. Three kinds of common mistakes emerged.

The top error involves the indirect object (iobj). All four occurrences in the corpus were misconstrued as direct objects.

The second kind of error was due to unawareness of an implicit copula verb. When a copula

exists or is implied, the label between the subject and predicate should be topic (top) rather than (nsubj); and the label between the subject and a noun should be attributive (attr) rather than direct object (dobj). Almost all mistakes with the labels top and attr fell into this category.

Third, as another technical mistake, students often failed to use the label preposition object (pobj), and substituted it with the more common direct object (dobj) instead.

Label	Error rate	Label	Error rate
iobj	100.0%	npadvmod	28.6%
attr	55.0%	nsubj	15.1%
top	50.0%	dobj	12.6%
pobj	35.0%	vmod	6.4%

Table 3. Dependency labels with the highest error rates.

5 Research Assignment

Combining the effort of the whole class, 176 of the 320 poems in the *Three Hundred Tang Poems*, comprising about 5000 characters, had been compiled in a treebank.

As a demonstration of the value of their annotations, a research assignment, with eight questions on various linguistic aspects of the poems, was designed. Before the release of the assignment, two preparatory steps were needed: the instructor edited the students' annotations into a gold version, and imported the gold version onto a web-based interface that allows searching for occurrences of specific dependency relations. The user may specify the relevant child and/or head word, or only their POS, and optionally also the dependency label.

Most questions in the assignment required searching for particular dependency relations and observing the word usage therein. For example, students were to find compound nouns where the head noun is modified by the characters "spring" or "autumn", two seasons that appear frequently in formulaic expressions to convey atmosphere (e.g., "wind in the spring", "moon in the autumn"). They were then to recognize the head nouns attested to be modified by both ("grass", "sun" and "light"). As another example, students were to identify all sentences where the usual SVO order had undergone word inversion, and comment on those words that were intentionally given the emphasis. Other questions addressed pivot constructions and onomatopoeia words. Average student performance on these questions ranges between 70% and 90%.

Perhaps the most challenging question was on the phenomenon of parallelism. Classical Chinese poems are read in pairs of two lines, or *couplets*. The two lines in a couplet are expected to have similar syntactic structures, yet the nature and extent of this "similarity" remained an open question. Taking 16 couplets from the treebank as samples, students were to explain any dissymmetry in the pairs of dependency trees, and point out the most frequent differences. About 50% of the students offered ideas similar to the conclusions of a larger-scale study (Lee & Kong, in submission), i.e., that certain sets of POS tags may be considered acceptable as parallel (e.g., numbers and adjectives), and that low-level syntactic structures need not be identical.

6 Discussions and Conclusions

We have described an undergraduate course on language information technology where students collectively created a treebank, then applied it in a research assignment.

This course design is demanding for the instructor, who must correct a substantial amount of annotations, under time pressure to produce a gold version for use in the assignment. Moreover, assignment questions may need to be adjusted, since the annotation results are not available beforehand. It is also demanding for students, who must master the dependency annotation scheme quickly.

The rewards of this design, however, are manifold for students, instructor and scholarship. First, annotation errors indicate areas where students' grasp of grammar is weak, and thus informative for language teachers. Second, some annotations reveal alternative syntactic interpretations, never thought of by the instructor, and can contribute to studies on syntactic ambiguities. Third, the resulting treebank can serve as a linguistic resource for all scholars.

Most significantly, the research assignment lets students reap the rewards of the new knowledge they had labored to create, providing a convincing demonstration of the practical value of treebanking. In future versions of the course, we hope to continue building this "cycle of contributing and learning" (Crane et al., 2012), where students learn to contribute new knowledge, and share it with others so they can collectively discover yet more knowledge.

Acknowledgments

This work was partially supported by a grant from the Early Career Scheme of the General Research Fund (#9041849) from the Research Grants Council of Hong Kong, and by a Teaching Development Grant (#6000349) from City University of Hong Kong.

References

David Bamman and Gregory Crane. 2010. Corpus Linguistics, Treebanks and the Reinvention of Philology. In *Informatik* 2010, pages 542-551.

Emily Bender. 2007. Combining Research and Pedagogy in the Development of a Crosslinguistic Grammar Resource. *Proc. GEAF Workshop*.

Kalina Bontcheva, Hamish Cunningham, Valentin Tablan, Diana Maynard, and Oana Hamza. 2002. Using GATE as an Environment for Teaching NLP. *Proc. Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Eckhard Bick. 2005. Grammar for Fun: IT-based Grammar Learning with VISL. In: Henriksen, Peter Juel (ed.), *CALL for the Nordic Languages*. pp.49-64.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. *Proc. 3rd Workshop on Syntax and Structure in Statistical Translation*.

Gregory Crane, Bridget Almas, Alison Babeu, Lisa Cerrato, Matthew Harrington, David Bamman, and Harry Diakoff. 2012. Student Researchers, Citizen Scholars and the Trillion Word Library. *Proc. 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*.

John Lee. 2012. A Classical Chinese Corpus with Nested Part-of-Speech Tags. *Proc. Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*.

John Lee and Yin Hei Kong. 2012. A Dependency Treebank of Classical Chinese Poems. *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

John Lee and Tak-sum Wong, 2012. Glimpses of Ancient China from Classical Chinese Poems. *Proc. 24th International Conference on Computational Linguistics (COLING)*.

Martin Volk, Sofia Gustafson-Capková, David Hagstrand, and Heli Uibo. 2005. Teaching Treebanking. In: *Nordisk Sprogteknologi. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Museum Tusculanums Forlag. Copenhagen. 2005.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer, 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering* 11:pp.207—238.