Towards Universal Dependencies for Learner Chinese

John Lee, Herman Leung, Keying Li
Department of Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk, leung.hm@gmail.com, keyingli3-c@my.cityu.edu.hk

Abstract

We propose an annotation scheme for learner Chinese in the Universal Dependencies (UD) framework. The scheme was adapted from a UD scheme for Mandarin Chinese to take interlanguage characteristics into account. We applied the scheme to a set of 100 sentences written by learners of Chinese as a foreign language, and we report inter-annotator agreement on syntactic annotation.

1 Introduction

A learner corpus consists of texts written by non-native speakers. Recent years have seen a rising number of learner corpora, many of which are error-tagged to support analysis of grammatical mistakes made by learners (Yannakoudakis et al., 2011; Dahlmeier et al., 2013; Lee et al., 2016b). In order to derive overuse and underuse statistics on syntactic structures, some corpus have also been part-of-speech (POS) tagged (Díaz-Negrillo et al., 2010; Reznicek et al., 2013), and syntactically analyzed (Ragheb and Dickinson, 2014; Berzak et al., 2016). These corpora are valuable as training data for robust parsing of learner texts (Geertzen et al., 2013; Rehbein et al., 2012; Napoles et al., 2016), and can also benefit a variety of downstream tasks, including grammatical error correction, learner proficiency identification, and language learning exercise generation.

While most annotation efforts have focused on learner English, a number of large learner Chinese corpora have also been compiled (Zhang, 2009; Wang et al., 2015; Lee et al., 2016a). However, POS analysis in these corpora has been limited to the erroneous words, and there has not yet been any attempt to annotate syntactic structures. This study presents the first attempt to annotate Chinese learner text in the Universal Dependencies (UD) framework. One advantage of UD is the potential for contrastive analysis, e.g., comparisons between a UD treebank of standard Chinese, a UD treebank of language X and portions of a UD treebank of learner Chinese produced by native speakers of X.

The rest of the paper is organized as follows. Section 2 reviews existing treebanks for learner texts. Section 3 describes the adaptation of a Mandarin Chinese UD scheme to account for non-canonical characteristics in learner text. Section 4 reports inter-annotator agreement.

2 Previous work

Two major treebanks for learner language — the Treebank of Learner English (TLE) (Berzak et al., 2016) and the project on Syntactically Annotating Learner Language of English (SALLE) (Ragheb and Dickinson, 2014) — contain English texts written by non-native speakers. TLE annotates a subset of sentences from the Cambridge FCE corpus (Yannakoudakis et al., 2011), while SALLE has been applied on essays written by university students. They both adapt annotation guidelines for standard English: TLE is based on the UD guidelines for standard English; SALLE is based on the POS tagset in the SUSANNE Corpus (Sampson, 1995) and dependency relations in CHILDES (Sagae et al., 2010).

Both treebanks adopt the principle of "literal annotation", i.e., to annotate according to a literal reading of the sentence, and to avoid considering its "intended" meaning or target hypothesis.

2.1 Lemma

SALLE allows an exception to "literal annotation" when dealing with lexical violations. When there is a spelling error (e.g., "*ballence"), the annotator puts the intended, or corrected form of the word ("balance") as lemma. For real-word spelling errors, the distinction between a word selection error and spelling error can be blurred. SALLE requires

Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 67–71,
Gothenburg, Sweden, 22 May 2017.

a spelling error to be "reasonable orthographic or phonetic changes" (Ragheb and Dickinson, 2013). For a sentence such as "... *loss its ballence", the lemma of the word "loss" would be considered to be "lose". The lemma forms the basis for further analysis in POS and dependencies.

To identify spelling errors, TLE follows the decision in the underlying error-annotated corpus (Nicholls, 2003). Further, when a word is mistakenly segmented into two (e.g., "*be cause"), it uses the UD relation `goeswith` to connect them.

2.2 POS tagging

For each word, SALLE annotates two POS tags, a "morphological tag" and a "distributional tag". The former takes into account "morphological evidence", i.e., the linguistic form of the word; the latter reflects its "distributional evidence", i.e., its syntactic use in the sentence. In a well-formed sentence, these two tags should agree; in learner text, however, there may be conflicts between the morphological evidence and the distributional evidence. Consider the word "see" in the sentence "*I have see the movie." The spelling of "see" provides morphological evidence to interpret it as base form (`VV0`). However, its word position, following the auxiliary "have", points towards a past participle (`VVN`). It is thus assigned the morphological tag `VV0` and the distributional tag `VVN`.

These two kinds of POS tags are similarly incorporated into a constituent treebank of learner English (Nagata et al., 2011; Nagata and Sakaguchi, 2016). They are also implicitly encoded in a POS tagset designed for Classical Chinese poems (Wang, 2003). This tagset includes, for example, "adjective used as verb", which can be understood as a morphological tag for adjective doubling as a distributional tag for verb. Consider the sentence 春風又綠江南岸 *chūnfēng yòu lǜ jiāngnán àn* "Spring wind again greens Yangtze's southern shore"[1]. The word *lǜ* 'green', normally an adjective, serves as a causative verb in this sentence. It is therefore tagged as "adjective used as a verb".

TLE also supplies similar information for spelling and word formation errors, but in a different format. Consider the phrase "a *disappoint unknown actor". On the one hand, the POS tag reflects the "intended" usage, and so "disappoint" is tagged as an adjective on the basis of its target hypothesis "disappointing". On the other hand, the "most common usage" of the original word, if different from the POS tag, is indicated in the TYPO field of the metadata; there, "disappoint" is marked as a verb.

[1] English translation taken from (Kao and Mei, 1971).

2.3 Dependency annotation

In both treebanks, "literal annotation" requires dependencies to describe the way the two words are apparently related, rather than the intended usage. For example, in the verb phrase "*ask you the money" (with "ask you *for* the money" as the target hypothesis), the word "money" is considered the direct object of "ask".

SALLE adds two new relations to handle non-canonical structures. First, when the morphological POS of two words do not usually participate in any relation, the special label '`-`' is used. Second, the relation `INCROOT` is used when an extraneous word apparently serves as a second root. In addition, SALLE also gives subcategorization information, indicating what the word can select for. This information complements distributional POS tags, enabling a comparison between the expected relations and those that are realized.

3 Proposed annotation scheme

Our proposed scheme for learner Chinese is based on a UD scheme for Mandarin Chinese (Leung et al., 2016). We adapt this scheme in terms of word segmentation (Section 3.1), POS tagging (Section 3.2) and dependency annotation (Section 3.3). We follow SALLE and TLE in adhering to the principle of "literal annotation", with some exceptions to be discussed below.

3.1 Word segmentation

There are no word boundaries in written Chinese; the first step of analysis is thus to perform word segmentation. "Literal annotation" demands an analysis "as if the sentence were as syntactically well-formed as it can be, possibly ignoring meaning" (Ragheb and Dickinson, 2014). As a rule of thumb, we avoid segmentations that yield non-existing words.

A rigid application of this rule, however, may result in difficult and unhelpful interpretations in the face of "spelling" errors. Consider the two possible segmentations for the string 不關 *bù guān* 'not concern' in Table 1. Literal segmentation should in principle be preferred, since *bù guān* are two words, not one. Given the context, however,

	Literal segmentation		Segmentation w/ spelling error
Text	不	關	不關
	bù	*guān*	*bùguān*
	'not'	'concern'	'not-concern'
Lemma	不	關	不管
	bù	*guān*	*bùguǎn*
	'not'	'concern'	'no matter'
POS	ADV	VERB	SCONJ

Table 1: Word segmentation of the string 不關 *bù guān* into two words (left) or one word (right), and the consequences on the lemma and POS tag.

the learner likely confused the character *guān* with the homophonous *guǎn*; the latter combines with *bù* to form one word, namely the subordinating conjunction 不管 *bùguǎn* 'no matter'. If so, the literal segmentation would misrepresent the semantic intention of the learner and yield an unhelpful syntactic analysis. We thus opt for the segmentation that assumes the spelling error; this interpretation, in turn, leads to *bùguǎn* as the lemma and SCONJ as the POS tag.

We follow SALLE in limiting spelling errors to orthographic or phonetic confusions. Specifically, for Chinese, the surface form and the lemma must have similar pronunciation[2] or appearance.[3]

3.2 POS tagging

Similar to SALLE, we consider both morphological and distributional evidence (Section 2.2). When non-native errors create conflicts between them, the former drives our decision on the POS tag, while the latter is acknowledged in a separate, "distributional" POS tag (henceforth, "POS_d tag"). In Figure 1, the POS tag for *kěpà* 'scary' is ADJ, reflecting its normal usage as an adjective; but its POS_d tag is VERB, since the pronoun *tā* 'him' suggests its use as a verb with a direct object.

The POS_d tag is useful for highlighting specific word selection errors involving misused POS (e.g., *kěpà* as a verb). It can also derive more general statistics, such as the use of adjectives where verbs are expected. In some cases, it suggests a target hypothesis (e.g., in Figure 1, to replace *kěpà* with

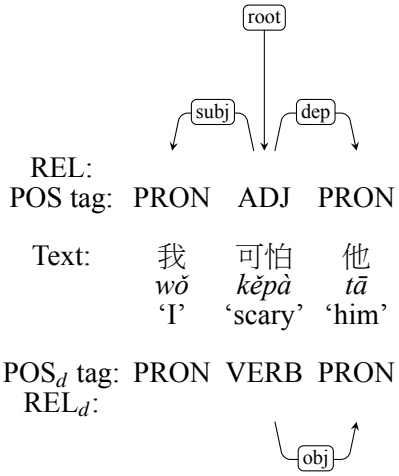

Figure 1: Parse tree for the sentence *wǒ kěpà ta* 'I scary him', likely intended as 'I scare him'. The POS tags and REL relations reflect the morphological evidence. Additionally, the POS_d tags (Section 3.2) and REL_d relations (Section 3.3) consider the distributional evidence.

a verb); but in others, a word insertion or deletion elsewhere might be preferred.

3.3 Dependency annotation

We now discuss how typical learner errors — word selection errors, extraneous words and missing words — may affect dependency annotation.

3.3.1 Word selection error

Dependency relation (henceforth, "REL") is determined on the basis of the POS tags rather than the POS_d tags. As long as these two agree, word selection errors should have no effect on dependency annotation. If a word's POS tag differs from POS_d, however, it can be difficult to characterize its grammatical relation with the rest of the sentence. In this case, we also annotate its "distributional relation" (henceforth, "REL_d") on the basis of its POS_d tag.[4]

Consider the sentence in Figure 1. From the point of view of POS tags, the relation between the adjective *kěpà* 'scary' and the pronoun *tā* 'him' is unclear. We thus assign the unspecified dependency, `dep`, as their REL.[5] From the point of view of POS_d tags, however, *kěpà* functions as a verb

[2] We allow different tones, such as {*guān*, *guǎn*}; and easily confusable pairs such as {j, zh} and {x, sh}.

[3] E.g., confusion between the characters 了 *le* and 子 *zǐ*.

[4] Similarly, Nagata and Sakaguchi (2016) use error nodes (e.g., `VP-ERR`) to annotate ungrammatical phrases (e.g., "*I busy").

[5] Similar to the underspecified tag '-' in SALLE, `dep` is used in English UD "when the system is unable to determine a more precise dependency relation", for example due to a "weird grammatical construction."

Agreement	Overall	Error span only
POS	94.0	91.0
POS_d	93.7	89.7
REL	82.8	75.1
REL_d	82.1	73.8

Table 2: The percentage of POS tags and labelled attachment on which the two annotators agree, measured overall and within text spans marked as erroneous.

and takes *tā* as a direct object, with the relation `obj` as their REL_d.

3.3.2 Extraneous words

When a word seems extraneous, we choose its head based on syntactic distribution. For example, the aspect marker 了 *le* must modify the verb that immediately precedes it with the relation 'auxiliary' (`aux`). Even when *le* is extraneous — i.e., when the verb should not take an aspect marker — we would annotate it in the same way.

A more difficult situation arises when there is no verb before the extraneous *le*, e.g., in the sentence * 我被了他打 *wǒ bèi le tā dǎ* 'I PASS ASP he hit' ("I was hit by him"). In this case, we choose *bèi* as head of *le* on account of word order, but the relation is `dep` rather than `aux`.

3.3.3 Missing words

When a word seems missing, we annotate according to UD guidelines on promotion by head elision. For example, in the sentence fragment 在中國最近幾年 *zài zhōngguó zuìjìn jǐ nían* 'in China recent few years', we promote *nían* 'year' to be the root. Although both *zhōngguó* 'China' and *nían* would be `obl` dependents if a verb was present, *nían* is promoted because it is closer to the expected location of the verb.

4 Evaluation

We harvested a 100-sentence evaluation dataset from the training data of the most recent shared task on Chinese grammatical diagnosis (Lee et al., 2016b). The dataset included 20 sentences with extraneous words, 20 with missing words, 20 with word-order errors, and 40 with word selection errors. In order to include challenging cases of word selection errors, i.e., those involving misuse of POS (Section 3.3.1), we examined the target hypothesis in the corpus. We selected 20 sentences where the replacement word has a different POS, and 20 sentences where it is the same. Two annotators, one of whom had access to the target hypothesis, independently annotated these sentences.

Word segmentation achieved 97.0% precision and 98.9% recall when one of the annotators was taken as gold. After reconciling their segmentation, each independently annotated POS tags and dependency relations. The inter-annotator agreement is reported in Table 2. Overall agreement is 94.0% for POS tags and 82.8% for REL (labeled attachment). The agreement levels are comparable to those reported in (Ragheb and Dickinson, 2013), where agreement on labeled attachment ranges from 73.6% to 88.7% depending on the text and annotator. One must bear in mind, however, that annotation agreement for standard Chinese is also generally lower than English.

Annotation agreement based on distributional evidence — i.e., POS_d and REL_d — is slightly lower. This is not unexpected, since it requires a higher degree of subjective interpretation. The most frequent discrepancies between morphological and distributional tags are ADJ vs. VERB, i.e., an adjective used as a verb (as in Figure 1); and VERB vs. NOUN, i.e. a verb used as a noun.

Annotation agreement is also lower within text spans marked as erroneous in the corpus, with agreement dropping to 91.0% for POS tags and 75.1% for labeled attachment. Further analysis revealed that agreement is especially challenging for word selection errors whose target hypothesis has a different POS. A post-hoc discussion among the annotators suggests that multiple plausible interpretations of an ungrammatical sentence was the main source of disagreement. For these cases, more specific guidelines are needed on which interpretation — e.g., considering a word as extraneous, as missing, or misused in terms of POS — entails the most literal reading.

5 Conclusions and Future Work

We have adapted existing UD guidelines for Mandarin Chinese to annotate learner Chinese texts. Our scheme characterizes the POS and dependency relations with respect to both morphological and distributional evidence. While the scheme adheres to the principle of "literal annotation", it also recognizes spelling errors when determining the lemma. Evaluation results suggest a reasonable level of annotator agreement.

Acknowledgments

This work is partially supported by a Strategic Research Grant (Project no. 7004494) from City University of Hong Kong.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. In *Proc. ACL*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications*.

Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proc. 31st Second Language Research Forum (SLRF)*.

Yu-Kung Kao and Tsu-Lin Mei. 1971. Syntax, Diction, and Imagery in T'ang Poetry. *Harvard Journal of Asiatic Studies*, 31:49–136.

Lung-Hao Lee, Li-Ping Chang, and Yuen-Hsien Tseng. 2016a. Developing Learner Corpus Annotation for Chinese Grammatical Errors. In *Proc. International Conference on Asian Language Processing (IALP)*.

Lung-Hao Lee, Gaoqi Rao, Liang-Chih Yu, Endong Xun, Baolin Zhang, and Li-Ping Chang. 2016b. Overview of NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. In *Proc. 3rd Workshop on Natural Language Processing Techniques for Educational Applications*.

Herman Leung, Rafaël Poiret, Tak sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. In *Proc. Workshop on Asian Language Resources*.

Ryo Nagata and Keisuke Sakaguchi. 2016. Phrase Structure Annotation and Parsing for Learner English. In *Proc. ACL*.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a Manually Error-tagged and Shallow-parsed Learner Corpus. In *Proc. ACL*.

Courtney Napoles, Aoife Cahill, and Nitin Madnani. 2016. The Effect of Multiple Grammatical Errors on Processing Non-Native Writing. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications*.

Diane Nicholls. 2003. The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT. In *Proc. Computational Linguistics Conference*.

Marwa Ragheb and Markus Dickinson. 2013. Inter-annotator Agreement for Dependency Annotation of Learner Language. In *Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications*.

Marwa Ragheb and Markus Dickinson. 2014. Developing a Corpus of Syntactically-Annotated Learner Language for English. In *Proc. 13th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Ines Rehbein, Hagen Hirschmann, Anke Lüdeling, and Marc Reznicek. 2012. Better tags give better trees — or do they? *LiLT*, 7(10):1–18.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture. In Ana Díaz-Negrillo, editor, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123, Amsterdam. John Benjamins.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2010. Morphosyntactic Annotation of CHILDES Transcripts. *Journal of Child Language*, 37(3):705–729.

Geoffrey Sampson. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford, UK.

Maolin Wang, Shervin Malmasi, and Mingxuan Huang. 2015. The Jinan Chinese Learner Corpus. In *Proc. 10th Workshop on Innovative Use of NLP for Building Educational Applications*.

Li Wang. 2003. *The metric of Chinese poems (Hanyu shiluxue 漢語詩律學)*. Zhonghua shuju, Hong Kong.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proc. ACL*.

Baolin Zhang. 2009. The Characteristics and Functions of the HSK Dynamic Composition Corpus. *International Chinese Language Education*, 4(11).