Is Argument Structure of Learner Chinese Understandable: A Corpus-Based Analysis

Yuguang Duan[♣], Zi Lin[♠♣] and Weiwei Sun[♠♡]

[♠]*Institute of Computer Science and Technology, Peking University*
[♠]*The MOE Key Laboratory of Computational Linguistics, Peking University*
[♣]*Department of Chinese Literature and Linguistics*
[♡]*Center for Chinese Linguistics, Peking University*
{ariaduan,zi.lin,ws}@pku.edu.cn

Abstract

This paper presents a corpus-based analysis of argument structure errors in learner Chinese. The data for analysis includes sentences produced by language learners as well as their corrections by native speakers. We couple the data with semantic role labeling annotations that are manually created by two senior students whose majors are both Applied Linguistics, The annotation procedure is guided by the Chinese PropBank specification, which is originally developed to cover first language phenomena. Nevertheless, we find that it is quite comprehensive for handling second language phenomena. The inter-annotator agreement is rather high, suggesting the understandability of learner texts to native speakers. Based on our annotations, we present a preliminary analysis of competence errors related to argument structure. In particular, speech errors related to word order, word selection, lack of proposition, and argument-adjunct confounding are discussed. We are also concerned with building interlanguage-specific natural language processing systems. We present a preliminary evaluation of three representative semantic role labeling models to gauge how successful a computational model can be to automatically analyze argument structures for learner Chinese.
Keywords: Learner Chinese, competence error, argument structure, semantic role labeling

1 Introduction

Corpus-based linguistic analysis is one of the fastest-growing methodologies in contemporary linguistics. It utilizes a large and principled collection of natural texts, known as a "corpus", as the basis for analyzing the actual patterns of use in natural texts. This method makes extensive use of computers for analysis, using both automatic and interactive techniques. It is important to both linguistic study and Natural Language Processing (NLP).

The corpus-based approach has been applied to studying not only the language patterns of native speakers but also the typical competence errors of foreign language learners. A learner corpus collects the language produced by people learning a foreign language, which have been essential for building NLP systems related to learner languages, as reported by (Nagata and Sakaguchi, 2016; Berzak et al., 2016). Furthermore, L2-L1[1] parallel treebanks have been shown beneficial for learner language analysis (Lee et al., 2017). On the one hand, learner languages exhibit general linguistic properties from part-of-speech and morphology, via phrase structure and/or dependency analysis, to aspects of meaning and discourse, function and style. On the other hand, learner languages exhibit cross-lingual influence-related properties, such as different types of learner *errors*, again ranging from the lexical and syntactic structures to discourse, function and usage.

Presently, most learner corpora aim to provide *grammatical error* labeling as well as correction. This can be helpful to second language teaching to a certain degree, but not comprehensive enough for more intelligent tasks, such as automatic essay scoring and determining the native language of a language learner. In such tasks, we not only need to spot the grammatical errors, but also need to scrutinize the logic and semantic propriety to evaluate the coherency of the discourse structure, the persuasiveness of the argumentation, etc. This requires a corpus with more exquisite syntactic and semantic labels, e.g., argument structure, that supports various Machine Learning algorithms for building NLP systems. This will also facilitate the probe to languages: statistical models can be employed to analyze learner corpora to provide insights into the nature of language acquisition or typical learner needs.

In this paper, we build an L2-L1 parallel corpus which has scrutinized semantic role labeling (SRL), and apply it to analysis on learner texts and certain automatic labeling tasks, in order to give a sample evaluation on the poten-

[1] In this paper, we call sentences written by non-native speakers (henceforth, "L2 sentences"), aligned to their corrections by native speakers (henceforth, "L1 sentences") L2-L1 parallel sentences.

tial of such resources. We present a preliminary analysis of competence errors related to argument structure. In particular, speech errors related to word order, word selection, lack of proposition, and argument-adjunct confounding are discussed. We also report a preliminary evaluation on three representative semantic role labeling models to gauge how successful a computational model can be to automatically analyze argument structures for learner Chinese. More detailed evaluations and analyses can found in the sister paper of this work (Lin et al., 2018).

2 Background

NLP is an area of computer science and artificial intelligence concerned with the automatic manipulation of natural language, like speech and text, by computers. In order to develop appropriate tools and techniques, NLP involves gathering of knowledge on how human beings understand and use language (**?**). NLP involves a wide range of tasks related to natural languages from basic to advanced.

SRL is a widely-studied NLP task that assigns semantic role labels to words or phrases in a sentence that indicate the argument structures. It consists of detecting the semantic arguments associated with the predicate of a sentence and assigning semantic roles to them according to their relationship to that predicate. Typical semantic roles include *Agent*, *Patient*, *Source*, *Goal*, and so forth, which are core arguments to a predicate, as well as *Location*, *Time*, *Manner*, *Cause*, and so on, which are adjuncts. See (1) for an example. In this sentence, 做 ("made") is the predicate, and is therefore labeled with rel. 我 ("I") is the agent of the predicate and 一个模型 ("a model") is the patient, so they are each labeled with $A0$ and $A1$. SRL is important to understand the essential meaning of the original input language sentences – *who* did *what* to *whom*, for *whom* or *what*, *how*, *where*, *when* and *why*, for it provides sentence-level semantic analysis of text that characterizes events.

(1) a. 我 做 了 一个 模型
I make. Past a model
'I made a model'

b. [我]$_{A0}$ [做]$_{rel}$ 了 [一个模型]$_{A1}$。

To build computation systems to automatically produce semantic role analysis, the state-of-the-art techniques leverage large-scale semantic role annotations, a collection of natural language sentences coupled with manually-assigned semantic structures. The Chinese PropBank (CPB; Xue and Palmer, 2009) is such a popular semantically annotated corpus for research on Chinese SRL. It adds a layer of predicate–argument structures to the Chinese TreeBank, assigning semantic role labels to syntactic constituents (rather than to the headwords in a dependency structure) in a sentence. Each verb has several *framesets* that are annotated with a fixed number of arguments: the core arguments of a predicate are labeled with a contiguous sequence of integers, in the form of *AN* (*N* is a natural number); the adjuncts are annotated with the label *AM* followed by a secondary tag that provides semantic information such as location, manner, and time. All the labels are defined by a general set of guidelines.

3 Semantic Role Labeling of An L2-L1 Parallel Corpus

Motivated by the importance of corpus in both (quantitative) linguistic analysis and building NLP systems, we are concerned with constructing a semantic role-annotated L2 corpus. To this end, we need to gather a L2 corpus in advance. In this paper, we use Lang-8, which contains large-scale learner texts of Mandarin Chinese that are collected from "language exchange" social networking services (SNS), a language-learning website where native speakers freely choose learners' essays to correct. The collecting work was done by our lab member Yuanyuan Zhao (Zhao et al., 2018), following (Mizumoto et al., 2011). By collecting the essays written by foreign learners and their revised version by Chinese natives, an initial corpus was set up consisting of 1,108,907 sentence pairs from 135,754 essays. As there is lots of noise in raw sentences, a series of measures are applied to cleaning up the data, and finally this learner corpus consists of 717,241 learner sentences from writers of 61 different native languages. As one sentence may be corrected by several correctors, we extracted **1,220,690** sentence pairs in total, including 310,075 sentences written by English native speakers and 484,140 b'y Japanese native speakers.

To make sure that a L2 corpus with semantic role labels is achievable, we first examined whether the learner texts can be understood by native speakers. To this end, we first conducted an inter-annotation between two annotators whose majors are Applied Linguistics to see if a high agreement can be achieved. In this process, we created a corpus consisting of manually-annotated predicate–argument labels on 600 L2-L1 pairs for learner Chinese.

We also notice that mother languages of language learners have a great impact on grammatical errors and hence influence the following ontological study and automatic semantic analysis. Therefore, our corpus includes four typologically different languages, i.e., English (ENG), Japanese (JPN), Russian (RUS) and Arabic (ARA). Each has a sub-corpus consisting of 150 sentence pairs.

Our annotators first annotated 50 parallel sentences for each native language, adapting PropBank specification as annotation heuristics, and then produced an initial adjudicated gold standard according to these 400 sentences. Based on this gold standard, the annotators proceeded to annotate a 100-sentence set for each language. The inter-annotator agreement is reported on these larger sets.

We also produced an adjudicated gold standard version of all 600 annotated sentences by comparing the annotation

of each annotator, discussing the differences, and either selecting one as fully correct or creating a hybrid representing the consensus decision for each choice point. When we felt that the decisions were not already fully guided by the existing annotation guidelines, we worked to articulate an extension to the guidelines that would support the decision.

This corpus can be utilized for both linguistic investigation and evaluation on NLP systems.

4 Semantic Analysis of L2-L1 Parallel Sentences

4.1 Inter-annotator Agreement

		P	R	F
ENG	L1	95.87	96.17	96.02
	L2	94.78	93.06	93.91
JPN	L1	97.95	98.69	98.32
	L2	96.07	97.48	96.77
RUS	L1	96.95	95.41	96.17
	L2	97.04	94.08	95.53
ARA	L1	96.95	97.76	97.35
	L2	97.12	97.56	97.34

Table 1: Inter-annotator agreement.

We calculate the precision (P), recall (R), F-score (F) to measure the inter-annotator agreement, as shown in Table 1. The inter-annotator agreement indicates that semantic annotations between the two annotators for both L1 and L2 sentences are quite consistent. All L1 texts have F-scores above 95, comparable to the annotation of CPB (Xue and Palmer, 2009). We take this result as a reflection that our annotators are qualified. F-scores of L2 sentences are all above 90, just a little bit lower than those of L1, indicating that L2 sentences can be greatly understood by native speakers.

		ENG	JPN	RUS	ARA
L1	*A0*	97.23	99.10	97.66	98.22
	A1	96.70	96.99	98.05	98.34
	A2	88.89	100.00	100.00	92.59
	A3	100.00	100.00	100.00	100.00
	A4	100.00	-	-	100.00
	AM	94.94	98.35	93.07	96.02
L2	*A0*	94.09	95.77	97.92	97.88
	A1	90.68	97.93	97.40	98.68
	A2	88.46	100.00	95.24	93.33
	A3	100.00	100.00	100.00	-
	A4	100.00	-	-	-
	AM	96.97	96.51	91.78	96.02

Table 2: Inter-annotator agreement (F-scores) relative to languages and role types.

Table 2 further reports agreements on each argument (*AN*) and adjunct (*AM*) in detail, according to which the high scores are attributed to the high agreement on arguments (*AN*). The labels of *A3* and *A4* have no disagreement since they are sparse in CPB and are usually used to label specific semantic roles that have little ambiguity.

4.2 Inter-annotator Disagreement

4.2.1 Disagreement on *A2* in Learner English

From Table 1 and 2, we notice that the F-score of English L2 (93.91) is relatively low compared to the other L2s for the low agreement on *A2* (88.46). We find that most L2 sentences with *A2* disagreements appear to have a mismatch between the Chinese and English attributive clause syntax. Take sentence in (1a) for an example.

(2) a. 我 帮 他们 盖 了 一个 盒子 可以 装满 泥土。
I help.PAST they.ACC make ASP a box can fill soil.
I helped them make a box that can be filled with soil.

b. 我 帮 他们 盖 了 一个 可以 装满 泥土 的 盒子。
I help.PAST they.ACC make ASP a can fill soil DE box.
I helped them make a box that can be filled with soil.

c. [我]$_{A0}$ [帮]$_{rel}$ [他们]$_{A1}$ [盖了一个盒子可以装满泥土]$_{A2}$。
[我]$_{A0}$ [帮]$_{rel}$ [他们]$_{A1}$ [盖了一个盒子]$_{A2}$ 可以装满泥土。

In the sentence, 装满泥土的盒子 (“a box that can be filled with soil”) should be treated as a whole and labeled with *A2* (thing *A0* helps *A1* with), as 装满泥土（“that can be filled with soil”）is the attributive clause of 盒子（“a box”）in English grammar. However, in Chinese, it should be written like (1b) where the attributive elements are put in front of the noun appended with an auxiliary word 的. This syntactic difference between the two languages causes the annotators to have splitting ideas on the boundaries of *A2*, as shown in (1c).

4.2.2 Disagreement on *AM* in all L2 Sentences

Another source of disagreement is the labels of *AM*. We analyze those L2 sentences with different annotated labels from two annotators, and find five predominant types of error, as described in Table 3.

Word order This error occurs when the sentence switches the position of constituents. In the example, 离开鄂木斯克 (“leave Emusike”) is the object of 打算 (“try to”), so 别 (“don’t”) should be labeled as a negative adjunct of 打算 (“try to”), while in the learner sentence 离开鄂木斯克 (“leave Emusike”) is transited forward, causing 别 (“don’t’) to become the adjunct of 离开 (“leave”) according to the principle of proximity. The unconformity between semantic relationship and syntactic structure can easily lead to disagreement, as shown in Table 3. This type of error also causes semantic ambiguity that impedes the understandability of learner texts.

Error type		Example	Disagreement	%
Word Order	**L2**	别 离开 鄂木斯克 打算! Don't leave Emusike try!	[别]$_{AM}$[离开 鄂木斯克]$_{A1}$[打算]$_{rel}$!	39%
	L1	别 打算 离开 鄂木斯克! Don't try leave Emusike!	[别 离开 鄂木斯克]$_{A1}$[打算]$_{rel}$!	
Word Selection	**L2**	我 被 召唤 为 帮 他们 翻译。 I pass summon for help they.acc translate.	[我]$_{A1}$ 被 [召唤]$_{rel}$ 为帮他们翻译。	27%
	L1	我 被 召唤 去 帮 他们 翻译。 I pass summon for help they.acc translate.	[我]$_{A1}$ 被 [召唤]$_{rel}$ [为帮他们翻译]$_{AM}$。	
Ambiguity	**L2**	我 和 妈妈 一起 住。 I with mom together live.	[我 和 妈妈]$_{A0}$ [一起]$_{AM}$[住]$_{rel}$。 [我]$_{A0}$[和 妈妈]$_{AM}$[一起]$_{AM}$[住]$_{rel}$。	16%
Lack of proposition	**L2**	我 昨天 见面 他们 了。 I yesterday meet.past they.acc asp.	[我]$_{A0}$[昨天]$_{AM}$[见面]$_{rel}$[他们]$_{AM}$ 了。	10%
	L1	我 昨天 和 他们 见面 了。 I yesterday with they.acc meet.past asp.	[我]$_{A0}$[昨天]$_{AM}$[见面]$_{rel}$ 他们 了。	
AM-AN confounding	**L2**	我 想 流利 在 日语。 I want to be fluent in Japanese.	[我]$_{A0}$ 想 [流利]$_{rel}$[在 日语]$_{AM}$。	8%
	L1	我 想 日语 流利。 I want my Japanese to be fluent.	我 想 [流利]$_{rel}$ 在 [日语]$_{A0}$。	

Table 3: Descriptions and percentages of *AM* error types

There are three main types of word order errors and Table 4 summarizes their distribution. Specifically, (1) Temporal sequence error (TS), where locative expressions, time expressions, beneficiaries or modifiers of verb are posited in the wrong places; (2) Verb and object reversal (OV), where the object is put in front of the verb; (3) Subject and verb reversal (VS), where the learners put the verb before the subject.

Type	ENG	JPN	RUS	ARA
TS	81%	73%	46%	42%
OV	0%	27%	31%	25%
VS	0%	0%	8%	16%
Other	19%	0%	15%	17%

Table 4: Overall performances of the syntax-based and neural syntax-agnostic SRL systems on the L1 and L2 data. "-p" means parser.

Interestingly, we notice from the table that different mother languages have different impact on the choice of word order in learner languages. Chinese is an analytic language and maintains a rather fixed word order, i.e., SVO, which is the same as English. On the contrary, Japanese is an SOV language, while Russian and Arabic are more flexible in word order since they are morphologically-rich languages which heavily leverage grammatical markers to indicate grammatical functions as well as semantic roles. Accordingly, learners whose native languages are Japanese, Russian and Arabic make much more OV errors since their mother tongues significantly influence their choice of word order. Furthermore, learners whose native languages are Russian and Arabic make more VS errors as these two languages can switch subject and verb flexibly. As for TS error, this is typical in all L2 texts since Chinese adverbial constituents have a rather complicated word order as described in (Jiang, 2009).

Word Selection This error occurs when the sentence has wrong words, redundant words or is lack of certain constituents. In the example in Table 3, the preposition 为 ("for") in the L2 sentence should be 去 ("to") which introduces the adjunct of purpose. The wrong word selection caused one annotator refuse to label the adjunct behind it.

Ambiguity This error occurs when some Chinese sentences per se can lead to ambiguity. Sometimes the disagreement of *AM* can be caused by the ambiguity of Chinese itself. In the example sentence, "和妈妈" can either be "and mom" in which case it will be part of coordinative agents as *A0* or "with mom" that serves as an adjunct (*AM*) of the predicate.

Lack of Preposition This error occurs when an *AM* requires a preposition while the sentence leaves it out. The most frequent cause for this error is verb subcategorization, e.g., mistaking interactive verb as non-interactive verb. In the example in Table 3, the interactive verb 见面 ("meet") usually has two parties as coordinative subjects (*A0*) linked by an auxiliary word 和 ("with"). However, learners often omit 和 ("with") and put the second *A0* behind the verb which is rather confusing to the annotators.

***AM-AN* Confounding** This error occurs when the sentence mistakes *AN* as *AM*. In the example, the word 流利

("fluent") is a predicate in Chinese whose only argument (*A0*) should be the language that is fluent. However, the learner mistook the person who can speak a certain fluent language as *A0* and put the language in a locative *AM*. In this case, the annotator cannot decide whether to label the language or the person as *A0*.

5 Evaluating Robustness of SRL systems

5.1 Three SRL Systems

The feasibility of reusing the annotation specification for L1 implies that we can reuse *standard* CPB data to train an SRL system to process learner texts. To evaluate the *robustness* of state-of-the-art SRL algorithms, we evaluate two representative SRL frameworks. One is a traditional syntax-based SRL system (Gildea and Jurafsky, 2000; Xue, 2008) In particular, we employ the system introduced in Feng et al. (2012). For constituent parsing, we use two parsers for comparison: one is the Berkeley parser[2] (Petrov et al., 2006), the other is a minimal span-based neural parserStern et al. (2017). On the Chinese TreeBank (CTB; Xue et al., 2005), it outperforms the Berkeley parser for in-domain test. We call the corresponding SRL systems as the **PCFGLA-parser-based** and **neural-parser-based** systems. The second SRL framework leverages an end-to-end neural model to implicitly capture local and non-local information (Zhou and Xu, 2015; He et al., 2017). Because all syntactic information (including POS tags) is excluded, we call this system the **neural syntax-agnostic** system.

5.2 Main Results

	PCFGLA-p		Neural-p		Neural syntax	
	L1	L2	L1	L2	L1	L2
Arg.-F	73.18	68.52	74.05	69.20	74.22	67.99
Adj.-F	72.28	70.77	73.73	72.39	73.92	70.08
all-F	72.87	69.28	73.94	70.30	74.12	68.71

Table 5: Overall performances of the syntax-based and neural syntax-agnostic SRL systems on the L1 and L2 data. "-p" means parser.

The overall performances of the three SRL systems on both L1 and L2 data are shown in Table **??**. For all systems, significant decreases in different mother languages can be consistently observed, highlighting the weakness of applying L1-sentence-trained system to process learner texts. Comparing the two syntax-based systems with the neural syntax-agnostic system, we find that the overall ΔF, which denotes the F-score drops from L1 to L2, is smaller in the syntax-based framework than in the syntax-agnostic system. On English, Japanese and Russian L2 sentences, the syntax-based system has better performances though sometimes works worse on the corresponding L1 sentences, i.e., the syntax-based systems are more robust when handling learner texts.

[2] code.google.com/p/berkeleyparser/

6 Conclusion

In this paper, we present an L2-L1 parallel corpus for SRL on learner Chinese texts. This is achievable since the learner Chinese texts are quite understandable. Such a corpus can be applied to analyzing error patterns in terms of argument structure as well as evaluating the performance of NLP systems. In our case study on learner Chinese texts by speakers of four typological mother tongues, the errors are mainly caused by the mismatch between the learners' mother language and Chinese grammar, including word order, word selection, lack of proposition and argument-adjunct confounding. Moreover the type of mother language also has an impact on the competence error, which is particularly obvious for word order error. As for automatic semantic parsing for interlanguages. We reveal two facts that are important towards a deeper analysis of learner languages: (1) the weakness of applying L1-sentence-trained systems to process learner texts, and (2) the importance of syntactic parsing to SRL for interlanguages. The finding may facilitate second language learning and teaching as well as improve the automatic semantic analysis system.

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 *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Minwei Feng, Weiwei Sun, and Hermann Ney. 2012. Semantic cohesion model for phrase-based SMT. In *Proceedings of COLING 2012*, pages 867–878, Mumbai, India. The COLING 2012 Organizing Committee.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what' s next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 473–483.

Wenying Jiang. 2009. Acquisition of word order in Chinese as a foreign language.

John Lee, Keying Li, and Herman Leung. 2017. L1-l2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy. Association for Computational Linguistics.

Zi Lin, Yuguang Duan, Yuanyuan Zhao, Weiwei Sun, and Xiaojun Wan. 2018. Semantic role labeling for learner chinese: the importance of syntactic parsing and l2-l1 parallel data. In *Proceedings of EMNLP*.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ryo Nagata and Keisuke Sakaguchi. 2016. Phrase structure annotation and parsing for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1837–1847, Berlin, Germany. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. *arXiv preprint arXiv:1705.03919*.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11:207–238.

Nianwen Xue. 2008. Labeling Chinese predicates with semantic roles. *Computational Linguistics*, 34:225–255.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese treebank. *Nat. Lang. Eng.*, 15:143–172.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the nlpcc 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing*, pages 439–445, Cham. Springer International Publishing.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1127–1137.