2018 International Conference on Bilingual Learning and Teaching

25-27 October 2018, OUHK Jockey Club Campus

Title: Is the Argument Structure of Learners' Chinese Understandable: A Corpus-based Analysis

Name(s): Yuguang Duan[1], Zi Lin[12] and Weiwei Sun[234]

Affiliation:
1. Department of Chinese Literature and Linguistics, Peking University, China
2. The MOE Key Laboratory of Computational Linguistics, Peking University, China
3. Institute of Computer Science and Technology, Peking University, China
4. Center for Chinese Linguistics, Peking University, China

Email: ariaduan@pku.edu.cn, zi.lin@pku.edu.cn, ws@pku.edu.cn

Biodata

Our laboratory focuses on natural language processing and language engineering (NLP). We develop computational resources, including semantic and syntactic parsers and annotated corpora.

Abstract

This paper presents a corpus-based analysis of errors in argument structure in learners' Chinese. The data for analysis include sentences produced by language learners, as well as their correction by native speakers. We couple the data with semantic role labelling annotations that were manually created by two senior students, both of whom were majors in Applied Linguistics. The annotation procedure was guided by the Chinese PropBank specification, which was originally developed to cover first language phenomena. Nevertheless, we have found that it is quite comprehensive for handling second language phenomena. The inter-annotator agreement was 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 prepositions, and argument-adjunct confounding are discussed.

Keywords:

learners' Chinese, competence error, argument structure, semantic role labelling

Highlights

- We have built an L2-L1 parallel corpus with semantic role labelling.
- We applied this corpus to analyze learner texts.
- Five typical errors related to argument structure are discussed.

Introduction

Corpus-based linguistic analysis is one of the fastest-growing methodologies in contemporary linguistics. It has been applied to studying not only the patterns of native speakers' language (henceforth, "L1"), but also the typical competence errors of foreign learners' language (henceforth, "L2").

A learner corpus collects the language produced by people learning a foreign language, which has been essential for building Natural Language Processing (NLP) systems related to learners' languages, as reported by Nagata and Sakaguchi (2016) and Berzak et al. (2016). Furthermore, L2-L1 parallel treebanks have been shown to be beneficial for analyzing learners' language (Lee, Li & Leung, 2017).

At present, most learner corpora aim to provide the labelling of grammatical errors as well as corrections. This can be helpful for second language teaching to a certain degree, but it is not comprehensive enough for more demanding 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 scrutinize the logic and semantic propriety to evaluate the coherency of the discourse structure, and the persuasiveness of the argumentation. This requires a corpus with more syntactic and semantic labels, e.g., argument structure which supports various Machine Learning algorithms for building NLP systems. This also facilitates the probe into 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 built an L2-L1 parallel corpus which has scrutinized semantic role labelling (SRL), and applied it to the analysis of learner texts. We present a preliminary analysis of competence errors related to argument structure. In particular, speech errors related to word order, word selection, lack of prepositions, and argument-adjunct confounding are discussed.

Literature Review/Background

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, for example, *Location*, *Time*, *Manner* and *Cause*, which are adjuncts.

SRL is important for understanding the essential meaning of the original input language sentences – *who* did *what* to *whom*, for *whom* or *what*, *how*, *where*, *when* and *why* – as it provides sentence-level semantic analysis of text that characterizes events.

The Chinese PropBank (CPB) (Xue & Palmer, 2009) is 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 labelled with a contiguous sequence of integers, in the form of *AN* (*N* is a natural number); and 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.

Methodology of the Study

Motivated by the importance of a corpus in both (quantitative) linguistic analysis and building NLP systems, we aimed to construct a semantic role-annotated L2 corpus. To this end, we had to gather an L2 corpus in advance. In this study, we used Lang-8, which contains large-scale learner texts in Mandarin Chinese that were 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, Jiang, Sun, & Wan, 2018), following Mizumoto, Komachi, Nagata and Matsumoto (2011).

To make sure that a L2 corpus with semantic role labels was achievable, we first examined whether the learner texts could be understood by native speakers. To this end, we first conducted an inter-annotation between two annotators whose majors are in Applied Linguistics to see if a high agreement could 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 noticed that the 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 included four typologically different languages, i.e., English (ENG), Japanese (JPN), Russian (RUS) and Arabic (ARA). Each had a sub-corpus consisting of 150 sentence pairs.

Our annotators first annotated 50 parallel sentences for each native language, adapting PropBank specification as an 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.

Findings

Table 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

We calculate the precision (P), recall (R), and 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 were quite consistent. All L1 texts had F-scores above 95, comparable to the annotation of CPB (Xue & Palmer, 2009). We take this result as a reflection that our annotators are qualified. The F-scores for L2 sentences were all above 90, just a little lower than those of L1, indicating that L2 sentences can be well understood by native speakers.

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

		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 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 had no disagreement since they were sparse in CPB and were usually used to label specific semantic roles that had little ambiguity.

Discussion and Implications

Disagreement on A2 in learners' English

From Tables 1 and 2, it can be seen 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 found that most L2 sentences with *A2* disagreements appeared to have a mismatch between the Chinese and English attributive clause syntax. Take the sentence in (1a) as an example.

(1) 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 labelled with A2 (the 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 caused the annotators to have different ideas on the boundaries of A2, as shown in (1c).

Disagreement on AM in all L2 sentences

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

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

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

Word order: This error occurs when the sentence switches the position of its 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 labelled 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.

Word selection: This error occurs when the sentence has wrong words, redundant words or lacks 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 to refuse to label the adjunct behind it.

Ambiguity: This error occurs when some Chinese sentences per se can lead to ambiguity. Sometimes, the disagreement on *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 would be part of coordinative agents as *A0* or "with mom" that serves as an adjunct (*AM*) of the predicate.

Lack of prepositions: This error occurs when an *AM* requires a preposition while the sentence leaves it out. The most frequent cause for this error is verb sub-categorization, e.g., mistaking an interactive verb as a 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 was 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 – the 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 could not decide whether to label the language or the person as *A0*.

Conclusion

In this paper, we present an L2-L1 parallel corpus for SRL on learner Chinese texts. This was achievable since the learner Chinese texts are quite understandable. Such a corpus can be applied for 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 were mainly caused by the mismatch between the learners' mother language and Chinese grammar, including word order, word selection, lack of prepositions and argument-adjunct confounding. We also applied this L2-L1 corpus to certain automatic labelling tasks to give a sample evaluation of the potential of such resources. More details can found in the sister paper of this work (Lin et al., 2018).

References

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K.S … Katz, B. (2016). Universal dependencies for learner English. Proceedings of the *54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 737–746). Berlin, Germany.

Lee, J., Li, K., & Leung, H. (2017). L1-l2 parallel dependency treebank as learner corpus. Proceedings of the *15th International Conference on Parsing Technologies*, (pp. 44–49). Pisa, Italy. Association for Computational Linguistics.

Lin, Z., Duan, Y., Zhao, Y., Sun, W., & Wan, X. (2018). Semantic role labeling for learner Chinese: The importance of syntactic parsing and l2-l1 parallel data. Proceedings of EMNLPY 2018 *Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium.

Mizumoto, T., Komachi, M., Nagata, M., & Matsumoto, Y. (2011). Mining revision log of language learning sns for automated Japanese error correction of second language learners. Proceedings of the *5th International Joint Conference on Natural Language Processing*, (pp.147–155). Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Nagata, R., & Sakaguchi, K. (2016). Phrase structure annotation and parsing for learner English. Proceedings of the *54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 1837–1847). Berlin, Germany

Xue, N., & Palmer, M. (2009). Adding semantic roles to the Chinese treebank. *Nat. Lang. Eng.*, *15*, 143–172.

Zhao,Y., Jiang, N., Sun, W., & Wan, X. (2018). Overview of the nlpcc 2018 shared task: Grammatical error correction. In *Natural language processing and Chinese computing* (pp. 439–445). Cham: Springer International Publishing.