Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

A Retrieval System for Interlanguage Analysis

Lung-Hao LEE [a], **Li-Ping CHANG** [b],
Bo-Shun LIAO[a], **Wan-Ling CHENG**[b] **& Yuen-Hsien TSENG**[a*]
[a]*Information Technology Center, National Taiwan Normal University, Taiwan*
[b]*Mandarin Training Center, National Taiwan Normal University, Taiwan*
*samtseng@ntnu.edu.tw

Abstract: In this paper, we describe the development of a retrieval system that is designed for analyzing the interlanguage. We adopt the annotated TOCFL learner corpus as the target to explore the language acquisition for leaners of learning Chinese as a foreign language. An illustrative scenario is presented to demonstrate the functionalities of implemented prototype system. This system can be deemed as a computer-assisted tool for contrastive interlanguage analysis research.

Keywords: second language acquisition, learner corpora, Mandarin Chinese

1. Introduction

Learner corpora: the Longman Learners' Corpus, the International Corpus of Learner English (ICLE) (Granger, 2003), and the Cambridge Learner Corpus (CLC) (Nicholls, 2003), to name but a few, are important collection of foreign language learners' linguistic production for research of second language acquisition and foreign language teaching (Granger, 2002). To make learner corpora to be more useful, they must be annotated using defined error types for automatic or manual analysis (Díaz-Negrillo & Fernández-Domínguez, 2006).

From the viewpoint of engineering, annotated learner corpora can be employed to develop specific Natural Language Processing (NLP) systems for educational applications. For instances, computer-assisted essay writing (Milton, 1998), spell error checking (Yu et al., 2014; Tseng et al., 2015), and grammatical error detection/correction (Chodorow and Leacock, 2000; Izumi et al., 2003; Lee et al., 2013, Ng et al., 2014; Yu et al., 2014; Lee et al., 2015). From linguistic perspectives, interlanguage is the type of linguistic system used by the second-/foreign- language learners who are in the process of learning a target language. Contrastive Interlanguage Analysis (CIA) is the main methodology that combines research areas of corpus linguistics and second language acquisition (Granger, 2015). Comparing learner corpus with native speaker's usages, researchers can identify learners' incorrectly linguistic usages or overgeneralized situations (Ishikawa, 2009). In addition, linguistic features of different L1s learners can also be obtained from CIA researches (Chang, 2014).

In this work, we develop and implement a retrieval system to help researchers to analyze the interlanguage. Our system is flexible to meet information needs in terms of various searching conditions. Besides, the search results can be downloaded easily if needed.

2. The Retrieval System of Annotated Leaner Corpus

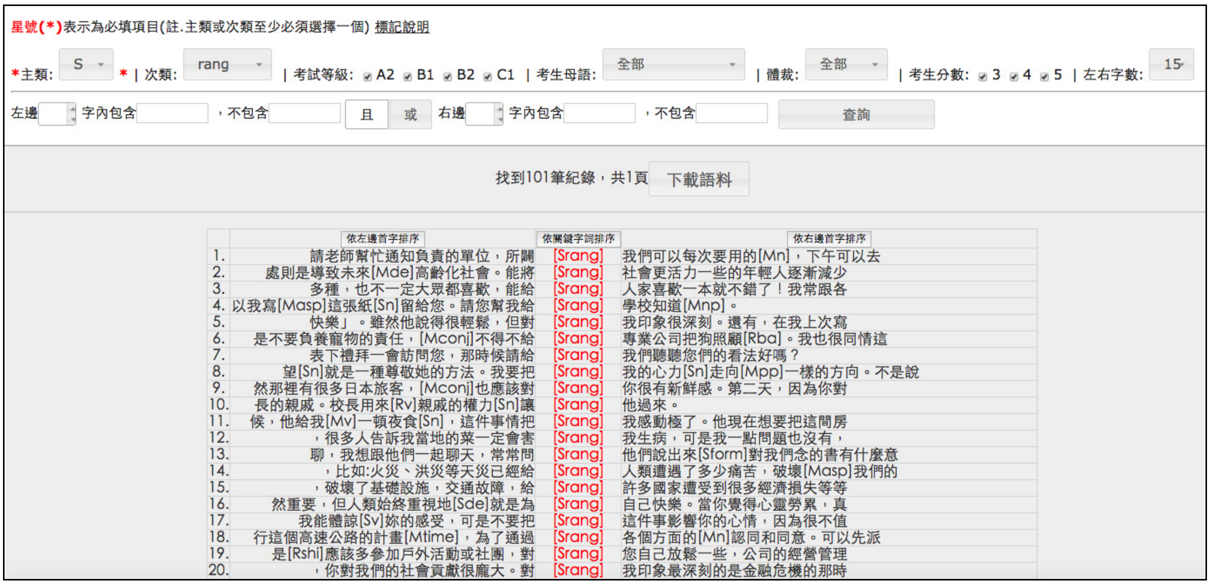
The learner corpus is mainly originated from the computer-based writing Test of Chinese as a Foreign Language (TOCFL). The writing test is designed according to the six proficiency levels of the Common European Framework of Reference (CEFR). Test takers have to complete two different tasks for each level. For example, for the A2 (Waystage level) candidates, they will be asked to write a note and describe a story after looking at four pictures. All candidates are asked to complete the writings on line. Each text is then scored on a 0-5 point scale. Score 5 means high-quality writings, score 3 is the threshold for passing the test, and so forth. There are 4,567 essays have been collected in the TOCFL learner corpus.

The native Chinese speakers are then trained and asked to label the grammatical error types of learners' writings using the tagging editor (Lee et al., 2014). For the purpose of studies in Chinese learners' interlanguage, hierarchical error tags are designed. One is target modification taxonomy, which includes mis-ordering (permutation), redundancy (addition), omission (deletion), and mis-selection (substitution). The other is linguistic category classification that consists of linguistic types, for example, noun, verb, preposition, specific construction, and so on. So far, 2837 essays with the score above 3 have been annotated. In total, there are 33, 497 error instances. The top 3 error tags are Sv (mis-Selection of verbs), Sn (mis-Selection of nouns), and Madv (Missing of adverbs). Their frequencies are 3838, 2252, 1714, respectively.

The searching functions of our retrieval system can be divided into two main parts: (1) *Basic search*: users can select the main types of error tags, i.e., modification types, and the linguistic categories. The levels of learner's language proficiency in CEFR and the scores of the learners' written essays can be chosen by ticking all that apply using checkbox. Searchers can also choose learner's mother-tongue language and types of writing styles. Besides, we also provide the concordance function to show the character contexts surrounding the search target in the search results. (2) *Advanced search*: when the search targets are determined, searchers can further filter the search results by including/excluding the characters occurring in the left-hand/right-hand sides. Moreover, the search results can be downloaded easily in plain text format for further research.

3. An Illustrative Scenario for Interlanguage Analysis

We present a scenario to illustrate the effectiveness of our developed retrieval system for interlanguage analysis. Take the '讓' (*rang4* 'to make') sentence for example, we can choose the main error type S and the sub-type *rang*. Figure 1 shows the searching results. We found that learners usually confuse '讓' (*rang4* 'to make') with '把' (*ba3* 'disposal marker'), '對' (*dui4* 'to someone'), and '給' (*gei3* 'to give'). If there is no error tag annotated in the corpus, even we search the keyword '讓' (*rang4* 'to make') and investigate one by one sentence to find the erroneous usages, but we cannot find the misused sentences with ba3 or dui4. With the help of this retrieval system, we can shorten the time efficiently. Moreover, we can limit the searching results into the specific word only, such as '把' (*ba3* 'disposal marker'), which will benefit to do deep observation and analysis. In addition to filtering function, we can also select the specific learners' attributes such as the learners' mother tongue or their proficiency. Take advantage of these functions, the analysis of interlanguages could be more easily and quickly done.



星號(*)表示為必填項目(註.主類或次類至少必須選擇一個) 標記說明

*主類: S * | 次類: rang | 考試等級: ☑ A2 ☑ B1 ☑ B2 ☑ C1 | 考生母語: 全部 | 體裁: 全部 | 考生分數: ☑ 3 ☑ 4 ☑ 5 | 左右字數: 15

左邊 字內包含 ，不包含 且 或 右邊 字內包含 ，不包含 查詢

找到101筆紀錄，共1頁 下載語料

	依左邊首字排序	依關鍵字詞排序	依右邊首字排序
1.	請老師幫忙通知負責的單位，所關	[Srang]	我們可以每次要用的[Mn]，下午可以去
2.	處則是導致未來[Mde]高齡化社會。能將	[Srang]	社會更活力一些的年輕人逐漸減少
3.	多種，也不一定大眾都喜歡，能給	[Srang]	人家喜歡一本就不錯了！我常跟各
4.	以我寫[Masp]這張紙[Sn]留給您。請您幫我給	[Srang]	學校知道[Mnp]。
5.	快樂」。雖然他說得很輕鬆，但對	[Srang]	我印象很深刻。還有，在我上次寫
6.	是不要負養寵物的責任，[Mconj]不得不給	[Srang]	專業公司把狗照顧[Rba]。我也很同情這
7.	表下禮拜一會訪問您，那時候請給	[Srang]	我們聽聽您們的看法好嗎？
8.	望[Sn]就是一種尊敬她的方法。我要把	[Srang]	我的心力[Sn]走向[Mpp]一樣的方向。不是說
9.	然那裡有很多日本旅客，[Mconj]也應該對	[Srang]	你很有新鮮感。第二天，因為你對
10.	長的親戚。校長用來[Rv]親戚的權力[Sn]讓	[Srang]	他過來。
11.	候，他給我[Mv]一頓夜食[Sn]，這件事情把	[Srang]	我感動極了。他現在想要把這間房
12.	，很多人告訴我當地的菜一定會害	[Srang]	我生病，可是我一點問題也沒有，
13.	聊，我想跟他們一起聊天，常常問	[Srang]	他們說出來[Sform]對我們念的書有什麼意
14.	，比如:火災、洪災等天災已經給	[Srang]	人類遭遇了多少痛苦，破壞[Masp]我們的
15.	，破壞了基礎設施，交通故障，給	[Srang]	許多國家遭受到很多經濟損失等等
16.	然重要，但人類始終重視地[Sde]就是為	[Srang]	自己快樂。當你覺得心靈勞累，真
17.	我能體諒[Sv]妳的感受，可是不要把	[Srang]	這件事影響你的心情，因為很不值
18.	行這個高速公路的計畫[Mtime]，為了通過	[Srang]	各個方面的[Mn]認同和同意。可以先派
19.	是[Rshi]應該多參加戶外活動或社團，對	[Srang]	您自己放鬆一些，公司的經營管理
20.	，你對我們的社會貢獻很龐大。對	[Srang]	我印象最深刻的是金融危機的那時

Figure 1. Searching results of rang4 sentence in the annotated TOCFL corpus

4. Conclusions and Future Work

This article describes our retrieval system that can be applied to analyze error types in annotated learner corpora. An illustrative scenario of this prototype system is presented for lnterlanguage analysis. We will further collect researchers' feedbacks and discuss with them to enhance its functions.

Acknowledgements

This research was partially supported by the Ministry of Science and Technology, under the grant MOST 103-2221-E-003-013-MY3, MOST 104-2911-I-003-301 and the "Aim for the Top University Project" and "Center of Learning Technology for Chinese" of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan.

References

Chang, L.-P. (2014). Salient linguistic features of Chinese learners with different L1s: a corpus-based study. *International Journal of Computational Linguistics and Chinese Language Processing*, 19(2), 53-72.

Chodorow, M., & Leacock, C. (2000). An unsupervised method for detecting grammatical errors. *Proceedings of NAACL'00* (pp. 140-147). Seattle, Washington: ACL Anthology.

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *RESLA*, 19, 83-102.

Granger, S. (2002). A bird's eye view of learner corpus research. In Granger, S., Huang, J. & Petch-Tyson, S. (eds.) *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam & Philadelphia: Benjamins, 3-33.

Granger, S. (2003). The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538-546.

Granger, S. (2015). Contrastive interlanguage analysis: a reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24.

Ishikawa, S. (2009). Phraseology overused underused by Japanese learners of English: a contrastive interlanguage analysis. *Phraseology, Corpus Linguistics and Lexicography*, 87-100

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., & Isahara, H. (2003). Automatic error detection in the Japanese learner's English spoken data. *Proceedings of ACL'03* (pp. 145-148), Sapporo, Japan: ACL Anthology.

Lee, L.-H., Chang, L.-P., Lee, K.-C., Tseng, Y.-H., & Chen, H.-H. (2013). Linguistic rules based Chinese error detection for second language learning. *Proceedings of ICCE'13* (pp. 27-29), Bali, Indonesia: Asia-Pacific Society for Computers in Education.

Lee, L.-H., Lee, K.-C., Chang, L.-P., Yu, L.-C., Tseng, Y.-H., & Chen, H.-H. (2014). A tagging editor for learner corpus annotation and error analysis. *Proceedings of ICCE'14* (pp. 806-808), Nara, Japan: Asia-Pacific Society for Computers in Education.

Lee, L.-H., Yu, L.-C., & Chang, L.-P. (2015). Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 1-6), Beijing, China: ACL Anthology.

Milton, J. (1998). WORDPILOT: enabling learners to navigate lexical universes. *Proceedings of the International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Hong Kong, China.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. *Proceedings of CoNLL'14* (pp. 1-14), Biltmore, Maryland: ACL Anthology.

Nicholls, D. (2003). The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. *Proceedings of CL'03* (pp. 572-581), Lancaster, UK.

Tseng, Y.-H., Lee, L.-H., Chang, L.-P., & Chen, H.-H. (2015). Introduction to SIGHAN 205 bake-off for Chinese spelling check. *Proceedings of SIGHAN'15* (pp. 32-37), Beijing, China: ACL Anthology.

Yu, L.-C., Lee, L.-H., Tseng, Y.-H., & Chen, H.-H. (2014). Overview of SIGHAN 2014 bake-off for Chinese spelling check. *Proceedings of CLP'14* (pp. 126-132), Wuhan, China: ACL Anthology.

Yu, L.-C., Lee, L.-H., & Chang, L.-P. (2014). Overview of grammatical error diagnosis for learning Chinese as a foreign language. *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 42-47), Nara, Japan: Asia-Pacific Society for Computers in Education.