Computational Linguistics and Chinese Language Processing
Vol. 20, No. 1, June 2015, pp. i-vi

Guest Editorial:
Special Issue on Chinese as a Foreign Language

Lung-Hao Lee[*], Liang-Chih Yu[+], and Li-Ping Chang[#]

Abstract

This introduction paper describes the research trends of Chinese as a second/foreign language along with related studies. We also overview the research papers included in this special issue. Finally, we conclude the findings and offer the suggestions.

Keywords: Computer-Assisted Language Learning, Second Language Acquistion, Leaner Corpora, Interlanguage, Mandarin Chinese.

1. Introduction

China's growing global influence has prompted a surge of interest in learning Chinese as a Foreign Language (CFL) and this trend is expected to continue. However, whereas many computer-assisted learning tools have been developed for learning English, support for CFL learners is relatively sparse, especially in terms of tools designed to automatically evaluate learners' responses. For example, while Microsoft Word has integrated robust English spelling and grammar checking functions for years, such tools for Chinese are still quite primitive. Another trend in demanding automated tools for CFL learners is accelerated by the recent progress is online learning technology and platforms, especially the so called MOOC (Massive Open Online Course) where a huge number learners can enroll in a course. The MOOC idea and platform not only make more people acquaint with online courses, but also demand automatic technology to handle the large volume of assignments and tests that are submitted by the enrolled learners.

[*] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
E-mail: lhlee@nlg.csie.ntu.edu.tw

[+] Department of Information Management & Innovation Center for Big Data and Digital Convergence Yuan Ze University, Taiwan
E-mail: lcyu@saturn.yzu.edu.tw

[#] Mandarin Training Center, National Taiwan Normal University, Taiwan
E-mail: lchang@ntnu.edu.tw
The author for Correspondence is Li-Ping Chang.

In contrast to the booming research developments for learning English as a foreign language, relatively few studies and tools are available for CFL learners. Chang (1995) proposed a three-step approach that uses the whole context within a sentence for spelling correction. Similar to Chang's approach, Zhang *et al*. (2000) presented an approximate word-matching algorithm to detect and correct Chinese spelling errors with the help of three edit operations: character substitution, insertion, and deletion. Ren *et al*. (2001) tried a hybrid approach that combines a rule-based method and a probability-based method to check Chinese spelling errors. Huang *et al*. (2007) proposed a learning model based on Chinese phonemic alphabet for spelling error check. Wu *et al*. (2010) proposed relative position and parse template language models to detect Chinese errors written by US learners using the NCKU corpus. Yu & Chen (2012) proposed a classifier to detect word-ordering errors in Chinese sentences from HSK learner corpus. Chang et al. (2012) proposed a penalized probabilistic First-Order Inductive Learning (pFOIL) algorithm, which integrates Inductive Logic Programming (ILP), First-Order Inductive Learning (FOIL), and a penalized log-likelihood function for error diagnosis. Lee *et al*. (2013) handcrafted a set of linguistic rules with syntactic information to detect grammatical errors. Lee *et al*. (2014) developed a sentence judgment system using both rule-based and n-gram statistical methods to detect grammatical errors in Chinese sentences.

In addition to research papers, several workshops and shared tasks focused on Chinese learning have been organized. For example, Chinese spelling check bakeoffs were organized in annual SIGHAN workshops, that is, the first one was held as part of the SIGHAN-7 in IJCNLP 2013 (Wu *et al*., 2013); the second version was held in CIPS-SIGHAN joint CLP-2014 conference (Yu *et al*., 2014); the third evaluation will be held in SIGHAN-8 as a ACL-IJCNLP 2015 workshop (Tseng *et al*., 2015). The research community has also organized a series of workshops on Natural Language Processing Techniques for Educational Applications (NLP-TEA) to give special attention to researches that have taken computer-assisted Asian language learning into consideration. The first NLP-TEA workshop was held in conjunction with ICCE-2014, accompanying with a shared task on Chinese as a Foreign Language was organized (Yu *et al*., 2014). The second NLP-TEA will be held as one of ACL-IJCNLP 2015 workshops with a Chinese Grammatical Error Diagnosis shared task (Lee *et al*., 2015). In summary, all of these academic activities increase the visibility of Chinese educational application research in the NLP community.

This special issue aims at general topics related to CFL research. Topics of interest include, but are not limited to as follows. From engineering perspectives, computer-assisted techniques for Chinese learning are important, such as spelling error check, grammatical error correction, sentence judgment systems, automated essay scoring, educational data mining, and so on. From linguistic perspectives, research areas include second language acquisition and interlanguage analysis by using learner corpora.

In the rest of this introduction paper, we describe the research paper included in this special issue in Section 2. Finally, we conclude the findings accompanying with suggestions in Section 3.

2. Content of Special Issue

This special issue consists of six research papers, which were reviewed and recommended by at least two experts. We briefly describe them as follows.

The first paper "HANSpeller: A Unified Framework for Chinese Spelling Correction" proposes a framework based on an extended Hidden Markov model and the ranker-based models, along with a rule-based model for Chinese spelling error detection and correction. CLP-2014 CSC datasets are adopted to demonstrate promising performance of their approach.

The second paper "A Study on Chinese Spelling Check Using Confusion Sets and N-gram Statistics" expands the coverage of confusion sets using Shuowen Jiezi and the Four-Corner codes. They also build a two-character confusion set. N-gram statistics are applied with the help of expanded and constructed confusion sets for Chinese spelling error checking. Experimental results show the approach improves the performance achieving by their previous system on SIGHAN 2013 CSC Bake-off.

The third paper "Automatically Detecting Syntactic Errors in Sentences Writing by Learners of Chinese as a Foreign Language" describes how to detect Chinese grammatical errors based on automatically-generated and manually-handcrafted rules. They propose a KNGED algorithm to identify syntactic errors written by CFL learners. NLP-TEA CFL datasets are used to show the effectiveness of their approach.

The fourth paper "Automatic Classification of the "De" Word Usage for Chinese as a Foreign Language" focuses on the usage of morphosyntactic particle "De". LEM 2 algorithm is adopted for deriving the rule set and then classifying the {的, 得, 地} based on induced rules for correct usages. The method achieves good performance on NLP-TEA CFL datasets.

The fifth paper "The Error Analysis of "*Le*" Based on Chinese Learner Written Corpus" analyzes the usage and the error types of "*Le*" made by English-native learners at the beginning and intermediate level based on NTNU learner corpus. The error types include redundancy and mis-selection of *le*1, *le*2 and *le*(1+2). Their findings show *le*1 is the most commonly spotted error type, and there is a large number of "*le*1" and "*le*(1+2)" redundant usages. In addition, pedagogical suggestions are also provided.

The sixth paper "Cross-Linguistic Error Types of Misused Chinese Based on Learners' Corpora" presents the construction of a learner corpus named 'Full Moon Corpus' and the tagging system for error annotation. The authors use comparative analysis method to observe the “*yi* ‘one’ + classifier” phrase by English-native learners and Japanese-native learners and discuss the reasons of ‘overuse’ and ‘underuse’ phenomenon.

3. Conclusions

This paper describes the present research trends of Chinese as a foreign/second language. All research papers included in this special issue are also introduced.

To improve the performance of NLP tools for Chinese learning by machine learning, collecting real learners’ erroneous sentences as much as possible is a challenging issue. The coverage of erroneous types is another. And tagging different corpora using the same format and tag set for learner corpus development is the other difficulty. The best strategy to deal with these problems may be to ally with research teams and to share collected linguistic resources.

Acknowledgments

We would like to thank all of the authors for their support, and our special thanks go to the anonymous reviewers who contributed their valuable wisdom and time to the research community.

References

Chang, C.-H. (1995). A new approach for automatic Chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, 278-283.

Chang, R.-Y., Wu, C.-H., & Prasetyo, P. K. (2012). Error diagnosis of Chinese sentences using inductive learning algorithm and decomposition-based testing mechanism. *ACM Transactions on Asian Language Information Processing*, *11*(1), Article 3 (30 pages).

Huang, C.-M., Wu, M.-C., & Chang, C.-C. (2007). Error detection and correction based on Chinese phonemic alphabet in Chinese text. In *Proceedings of the 4th Conference on Modeling Decisions for Artificial Intelligence*, 463-476.

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. In *Work-in-Progress Poster Proceedings of 21st International Conference on Computers in Education*, 27-29.

Lee, L.-H., Yu, L.-C., & Chang, L.-P. (2015). Overview of shared task on Chinese grammatical error diagnosis. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications.*

Lee, L.-H., Yu, L.-C., Lee, K.-C., Tseng, Y.-H., Chang, L.-P., & Chen, H.-H. (2014). A sentence judgment for grammatical error detection. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations*, 67-70.

Ren, F., Shi, H., & Zhou, Q. (2001). A hybrid approach to automatic Chinese text checking and error correction. In *Proceedings of 2001 IEEE International Conference on Systems, Man, and Cybernetics*, 1693-1698.

Tseng, Y.-H., Lee, L.-H., Chang, L.-P., & Chen, H.-H. (2015). Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing*.

Wu, C.-H., Liu, C.-H., Harris, M. & Yu, L.-C. (2010). Sentence correction incorporating relative position and parse template language model. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1170-1181.

Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese spelling check evaluation at SIGHAN bake-off 2013. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*, 35-42.

Yu, C.-H., & Chen, H.-H. (2012). Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. In *Proceedings of the 24th International Conference on Computational Linguistics*, 3003-3018.

Yu, L.-C., Lee, L.-H., & Chang, L.-P. (2014). Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications*, 42-47.

Yu, L.-C., Lee, L.-H., Tseng, Y.-H., & Chen, H.-H. (2014). Overview of SIGHAN 2014 bake-off for Chinese spelling check. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 126-132.

Zhang, L., Huang, C., Zhou, M., & Pan, H.-H. (2000). Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 248-254.