A hybrid approach for correcting grammatical errors

Kiyoung Lee[1], Oh-Woog Kwon[2], Young-Kil Kim[3], and Yunkeun Lee[4]

Abstract. This paper presents a hybrid approach for correcting grammatical errors in the sentences uttered by Korean learners of English. The error correction system plays an important role in GenieTutor, which is a dialogue-based English learning system designed to teach English to Korean students. During the talk with GenieTutor, grammatical error feedback and better expressions are offered to learners. We surveyed the grammatical mistakes that occurred in the English sentences uttered by Korean learners. These errors involve preposition errors, verb form errors, agreement errors, noun countability errors and determiner errors. The hybrid error correction system consists of 5 components: an error memory based correction system, a machine learning based correction system, an n-gram based correction system, an edit distance based correction system and a selector. The correction performance of each component is different depending on error types. To evaluate the hybrid system, we used a test set comprising of 858 sentences extracted from utterances by Korean learners. The test set includes not only ungrammatical sentences, but also correct sentences. We conducted various experiments and examined the effect of the hybrid approach on grammatical error correction. The experiments show promising results for correcting grammatical errors.

Keywords: grammatical error correction, dialogue-based computer assisted language learning.

1. Electronics and Telecommunications Research Institute, Daejeon, Korea; leeky@etri.re.kr
2. Electronics and Telecommunications Research Institute, Daejeon, Korea; ohwoog@etri.re.kr
3. Electronics and Telecommunications Research Institute, Daejeon, Korea; kimyk@etri.re.kr
4. Electronics and Telecommunications Research Institute, Daejeon, Korea; yklee@etri.re.kr

How to cite this article: Lee, K., Kwon, O.-W., Kim, Y.-K., & Lee, Y. (2015). A hybrid approach for correcting grammatical errors. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 362-367). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000359

1. Introduction

Recently, there has been a growing interest in Computer-Assisted Language Learning (CALL). Particularly in Korea, the time and cost to learn English are enormous and have been on the rise every year. We have developed GenieTutor (Kwon et al, 2015), which is a dialogue-based English learning system for Korean learners. The system consists of non-native optimized speech recognition modules and semantic/grammar correctness evaluation based tutoring modules (Kwon, Lee, Kim, & Lee, 2015). A learner has a talk with GenieTutor on various topics of scenarios consisting of 3 to 4 turns. During the talk with GenieTutor, grammatical error feedback and better expressions are offered to learners. These scenarios help learners not to be out of basic flow of dialogue. Learners take lessons on pronunciation, grammar and useful expressions from conversation with the virtual tutor. In this paper, we describe the hybrid grammatical correction system. Section 2 of this paper gives an overview of our system to detect and correct grammatical mistakes. Section 3 illustrates experimental results. In section 4, we sum up the discussion and show the future research direction.

2. Method

2.1. Grammatical error types

The grammatical error correction system plays an important role in GenieTutor, which is a dialogue based English learning system. The task of the grammatical error correction system is to detect and to correct grammatical mistakes made by an English learner. We defined target errors based on the Cambridge Learner Corpus (Nicholls, 2003) and the NUS Corpus (Dahlmeier, Ng, & Wu, 2013). These errors frequently occur in sentences or utterances by Korean learners. Table 1 shows the grammatical error types which we aim to correct.

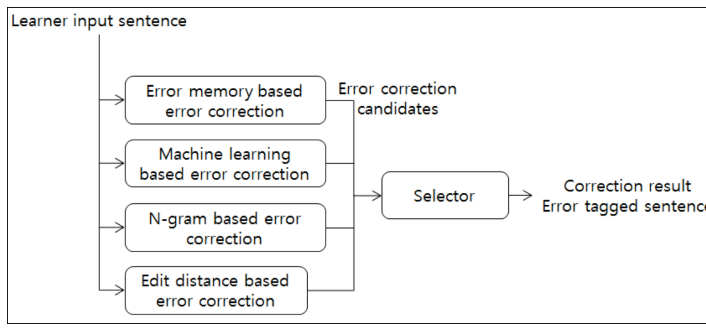
Table 1. Grammatical error types

Error Tag	Error Category	Error Tag	Error Category
RV	Replacing a verb	TV	Verb tense
FV	Verb form	AGV	Subject-verb agreement
MV	Missing a verb	UV	Unnecessary verb
RT	Replacing a preposition	MT	Missing a preposition
UT	Unnecessary preposition	MD	Missing a determiner
UD	Unnecessary determiner	RD	Replacing a determiner
RN	Replacing a noun	AGN	Noun agreement
FN	Noun form	MN	Missing a noun
UN	Unnecessary noun	CN	Noun countability

2.2. Hybrid grammatical error correction

Grammatical errors have their unique characteristics. The clues to detect and correct mistakes are also different from error types. There are various approaches to detect and correct grammar mistakes. We devised a hybrid correction system that combines four types of correction systems and a selector. Figure 1 shows the configuration of our hybrid grammatical error correction system. Each correction system takes as an input a sentence uttered by a learner and generates correction candidates according to their strategy to detect and correct errors. A selector then decides a final error type and a correction among correction candidates from each system.

Figure 1. The configuration of a hybrid grammatical error correction system



The knowledge for correction used by each component is based on 21,400 learner utterances excluding system's utterances in predefined scenarios.

2.2.1. Error memory based error correction

An error memory is a pattern with context and correction information. An error memory is as follows: *i am interest/interested/FV in music*.

The above error memory is applied to an input sentence "I am interest in music". In this sentence, 'interest' should be replaced by 'interested' and a mistake type is FV (Verb Form). The recall of error memory based correction is low. Its precision, however, is very high.

2.2.2. Machine learning based error correction

A machine learning based error correction system requires a grammatical error tagged training corpus for training classifiers. To build the training corpus we

automatically generated erroneous sentences and tagged error codes for the 21,400 sentences mentioned above. At the same time, we collected learner utterances from test service. Then, mistakes were tagged by human using error tags. We used a SVM classifier to detect and correct grammatical errors.

2.2.3. N-gram based error correction

N-gram data is extracted from common 21,400 sentences. In the n-gram correction model, the window size is set to 2 to 5 words. By replacing an input word with a possible form of the word, an n-gram model generates correction candidates and calculates their frequencies.

2.2.4. Edit distance based error correction

Dialogue scenarios consist of system utterances and corresponding correct answers. So, by searching correct answers that are most similar to an input sentence made by a learner, a correction candidate can be generated from the difference. An edit distance based error correction uses this characteristic.

2.2.5. Selector

A selector decides a final error type and correction information using the weight based on the performance of each correction system on error types. We assigned a different weight on each correction method depending on error types and their performance. For example, it is difficult for the error memory based correction system to find the mistakes detected by considering a broad context.

3. Experiments

To evaluate the hybrid grammatical error correction system, 858 sentences were randomly extracted from sentences uttered by Korean learners. The test set includes not only sentences involving words or phrases used ungrammatically, but also correct sentences. Table 2 and Table 3 show the precision and the recall on test.

Table 2. The performance comparison

System	Precision	Recall
Error memory based correction	98.6%	30.5%
Edit distance based correction	90%	15.5%
Machine learning based correction	64.2%	14.6%
N-gram based correction	65.3%	27.5%
Hybrid correction	**91.3%**	**45.1%**

Table 3. The performance of hybrid error correction

Error type	Precision	Recall
AGV	100%	68.8%
CN	60%	27.3%
FV	100%	50%
MD	92.1%	63.6%
MT	88.9%	29.6%
MV	100%	40%
RD	100%	70%
RT	82.4%	58.3%
UD	90%	75%
UT	100%	44.4%
UV	100%	25%
Total	**91.3%**	**45.1%**

We surveyed the effect of a hybrid grammatical error correction method. There still remain some problems to be solved:

- How to improve the performance of a hybrid error correction system for more general domains? Our system works well for dialogues similar to given scenarios. However, it is susceptible to correct grammatical errors in sentences which are out of given scenarios. We think that a machine learning based method and an n-gram method will be helpful to solve these coverage problems. So we are consistently collecting and building a training corpus and a n-gram data.

- In our hybrid system, it is very effective to maximize the performance of a selector. Modelling a selector by the performance of each correction system according to error types is needed.

- Because a false alarm is very critical for learning systems, we focused on correction precision for test service. By the same token, we assigned higher weight on the correction candidate of an error memory based system and an edit distance based system. As a future research direction, we consider to improve the recall of our hybrid method.

4. Conclusions

We have described a hybrid grammatical mistake correction system. Our hybrid error correction system consists of five components: an error memory based correction system, a machine learning based correction system, an n-gram based

correction system, an edit distance based correction system and a selector. Since grammatical errors are very diverse and have unique characteristics, it is difficult to cover these errors using only one correction system.

We plan to improve recall rate of our system on out of scenario sentences. To do that, the role of a machine learning and an n-gram based error correction approach is very important.

5. Acknowledgements

This work was supported by the ICT R&D program of MSIP/IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning]

References

Kwon, O.-W., Lee, K., Kim, Y.-K, & Lee, Y. (2015). GenieTutor: a computer assisted second-language learning system based on semantic and grammar correctness evaluations In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 330-335). Dublin Ireland: Research-publishing.net. doi:10.14705/rpnet.2015.000354

Kwon, , O.-W., Lee, K., Roh, H.-H., Huang, J.-X., Choi, S.-K., Kim, Y.-K, Jeon, H. B., Oh, Y. R., Lee, Y.-K., Kang, B. O., Chung, E., Park, J. G., & Lee, Y. (2015). GenieTutor: a computer assisted second-language learning system based on spoken language understanding, *Proceedings of the International Workshop on Spoken Dialog Systems*. Retrieved from https://www.uni-ulm.de/in/iwsds2015/list-of-accepted-papers.html

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: the NUS corpus of learner English. *Proceeding of the Eighth Workshop on Innovative Use of NLP for Building Educational Application* (pp. 22-31).

Nicholls, D. (2003). The Cambridge learner corpus – error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics* (pp.572-581).

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2015.