The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

Automatic Extraction of English Grammar Pattern Correction Rules

Kuan-Yu, Shen
Institute of Information Systems and Applications
National Tsing Hua University
kyshen@nlplab.cc

Yi-Chien, Lin
Department of Foreign Languages and Literature
National Tsing Hua University
nicalin@nlplab.cc

Jason S. Chang
Department of Computer Science
National Tsing Hua University
jason@nlplab.cc

Abstract

We introduce a method for creating error-correction rules for grammar pattern errors in a given annotated learner corpus. In our approach, annotated edits in the learner corpus are converted into edit rules for correcting common writing errors. The method involves automatic extraction of grammar patterns, and automatic alignment of the erroneous patterns and correct patterns. At run-time, grammar patterns are extracted from the grammatically correct sentences, and correction rules are retrieved by aligning the extracted grammar patterns with the erroneous patterns. Using the proposed method, we generate 1,499 high-quality correction rules related to 232 headwords. The correction rules are tested and about 36% of the essays are improved by applying these rules. The method can be used to assist ESL students in avoiding grammatical errors, and aid teachers in correcting students' essays. Additionally, the method can be used in the compilation of collocation error dictionaries and the construction of grammar error correction systems.

Keywords: grammar patterns, edit rules, pattern alignment, pattern extraction

1 Introduction

The importance of using correct grammar patterns is directly associated with the proficiency level of a language learner. English proficiency tests aiming at learners of English as a second language (ESL) such as TOEIC and TOEFL both include questions that require the examinee to have accurate knowledge on grammar patterns.

However, grammar patterns pose a great barrier to ESL learners for its inconsistency. For instance, "*talk about an issue*" is a grammatically correct phrase, while "*discuss about an issue*" is grammatically incorrect and should be corrected into "discuss an issue" . Hence, researches on the detection and correction of grammar pattern errors have been conducted with computational approaches.

Grammar patterns are rules that describe how words are used. A grammar pattern tells us the correct combination of a clause or phrase with a given verb, noun, or adjective. For instance, the verb discuss could be used with a prepositional phrase with with (*discuss with the manager*) or with a one noun phrase (*discuss the issue*).

Our goal is to convert the erroneous sentence into edit rules in the form of part of speech tags. For the example above, the correction of "*discuss about an issue*" to "*discuss an issue*" would be express as "*V about n* $\rightarrow$ *V n*" . By leveraging an annotated learner corpus Education First - Cambridge Open Language Database (EFCAMDAT) Geertzen et al. (2013) Huang et al. (2018), we retrieve the editing process and the frequency of the error being made. We automatically extract the

grammar patterns from the sentences, which patterns are provided by Collins Dictionary of Grammar Patterns.

This paper focuses on the algorithm of converting sentences with verb grammar pattern errors into edit rules. Our method successfully extracts 1,499 common grammar pattern errors over 232 headwords with a basic threshold of frequency above 10.

Comparing to the corpus of Longman Dictionary of Common Errors (Turton and Heaton, 1996), our result specifically focuses on the errors of grammar pattern and express the correction rules in a cleaner format. The correction rules constructed by our method are being experimented in a situation mimicking a teacher correcting a learner's essay. The result shows that about 36% of the essays contain errors that could be corrected by the rules.

The remaining of this paper would be organized as follows: Section 2 gives a background of previous works related to grammar pattern error correction. Section 3 presents our proposed method and the corpus used. Section 4 shows our experimental results and evaluation among other methods. Finally, section 5 provides a conclusion and insights for future studies.

2 Related Works

Grammatical error correction is an extensively studied topic. Numerous works have been conducted through rule-based and statistical approaches, specifically focus on the correction of prepositional errors written by ESL learners.

For rule-based approaches, Eeg-Olofsson and Knutsson (2003) defines a set of rules for detecting word, phrase, and prepositional errors in Swedish text. Bender et al. (2004) develops strategies regarding syntactic rules to reconstruct erroneous sentences into correct sentences. These approaches rely heavily on designed rules which require the time and labor of linguistic experts.

Statistical methods have been widely used due to the emerging of large text databases. Researchers apply statistical methods to correct prepositional errors in articles written by ESL learners. Sun et al. (2007) builds a classifier to identify erroneous and correct sentences. The features of the classifier, Labeled Sequential Patterns, are common patterns that indicate the errors or correctness of a sentence, which are closely related to grammar patterns. Brockett et al. (2006) uses phrasal Statistical Machine Translation (SAT) techniques to identify and correct writing errors made by learners. The proposed model maps small phrasal *"treelets"* generated by dependency parsing to grammatically correct strings, allowing the input erroneous sentence to be slightly ungrammatical, which is a typical feature of ESL learners.

Combining rule-based and statistical approaches, Chodorow et al. (2007) combines maximum entropy classifier and rule-based filters to detect preposition errors of student essays. The classifier is trained with contextual features regarding the Part-Of-Speech tags adjacent to the prepositions.

Recently, Huang et al. (2010) describes a framework to extract correction rules by calculating Levenshtein distance between correct and erroneous sentences. The framework is language independent and does not take linguistic features into account. Chen et al. (2017) considers grammar pattern and the semantic category of noun phrases while extracting the correction rules, establishing a writing suggestion system for language learners.

3 Method

Since we are interested in the edit rules for grammar pattern errors, we utilize a learner corpus with annotated edit process EFCAMDAT. The annotations of the corpus are corrections made by English experts. The corpus provides 2,300,000 sentences with annotations. We obtain the original sentences and the corrected sentences from the corpus, in which the former are assumed to be grammatically incorrect, and the latter to be grammatically correct. Since we are interested in grammar pattern errors, among all the edit tags that show the error type, we reserve only the ones with *XC (change of word)*, *D (deletion of word)*, *IS (insertion of word)*, *MW (missing of word)*, *PR (prepositional error)*, or *WC (word choice error)* tags. These tags are chosen for they are more relevant to grammar patterns (Geertzen et al., 2013) (Huang et al., 2018).

Our method could then be divided into two parts: Grammar pattern extraction and optimal alignment. After achieving the pairwise edit rules, a threshold could be set to improve the quality.

3.1 Extracting grammar patterns

Our grammar pattern data are taken from Collins COBUILD of Grammar Patterns[1], which provides up to 145 grammar patterns for verb. Collins COBUILD of Grammar Patterns is based on corpus research carried out by lexicographers, which lists all the grammar patterns used in English, and all the words regularly used with a given pattern.

We extract grammar patterns from the grammatically correct sentences. First, we merge each noun phrase into one single token by constituency parsing, and then perform part-of-speech tagging on all the tokens. We convert the tags of the tokens into a simplified form that adapts to our grammar pattern data. The conversion rules are manually written to adapt to the tool used for part-of-speech tagging and the grammar pattern data, which, in our experiment, SpaCy and Collins COBUILD of Grammar Patterns are used.

Grammar patterns are detected by sequence matching. The tokens of the grammatically correct sentence are iterated, and multiple patterns could be detected in a single phrase. An example of the whole process of grammar pattern extraction is provided (Table 1).

3.2 Aligning original and edited patterns

Since we had the edited grammar patterns of a given sentence, we then need to retrieve its unedited form to obtain the common grammar pattern errors. We use a dynamic programming approach *pairwise sequence alignment* to retrieve the unedited forms.

In pairwise sequence alignment algorithm, two sequences are aligned with the least cost (or highest score). In our approach, the first sequence is the extracted grammar pattern, while the other is a 5-gram phrase extracted from the unedited sentence, starting from the location of the grammar pattern's headword.

[1] https://grammar.collinsdictionary.com/grammar-pattern

Three conditions occurred in pairwise sequence alignment algorithm: gap, mismatch, and match. A gap indicates that a token of a sequence does not align to any token of another sequence. A mismatch indicates that a token of a sequence does align to a token of another sequence, but the two tokens are not identical. A match indicates that a token of a sequence is aligned to a token of another sequence, and the two tokens are identical. In our approach, gaps or mismatches acquire no score, and a match acquires 1 score. Pairwise sequences with scores below 2 are discarded.

After alignment, to ensure only one editorial occurred in a pairwise rule, tokens of both sequences are iterated simultaneously. This time, gaps or mismatches acquires -1 score, and a match acquires 1 score. We retrieve the pairwise pair with the highest score at the maximum length possible.

4 Results & evaluation

Using EFCAMDAT and Collins COBUILD as our reference data, our method successfully achieves 1,499 correction rules over 232 headwords with the basic threshold of frequency above 10. The usage of grammar pattern ensures the extracted patterns to be correct and meaningful. Threshold and reference data could be adjusted as needed. Table 2 shows part of our result.

Edit rules achieved from our method are pairwise, consist of common grammar pattern errors and their corrections. Our result clearly gives the headword, frequency, and examples of the edit rules.

Comparing to the Longman Dictionary of Common Errors, the rules are more explicit and concise and with much more examples, which could be utilized conveniently for further research and applications. Additionally, our result focuses on grammar pattern errors, while the Longman Dictionary of Common Errors covers all sorts of common errors, including word choice, spelling errors, and tense errors.

We examine the correction rules by providing the rules as suggestions for the corrector while correcting the essays written by ESL learners. The essays are provided by the ETS Corpus of Non-Native Written English, Lin-

Original sentence	Give	the elegant present	to	Tom	.
Merging noun-phrase	Give	<NP>	to	<NP>	.
POS tagging	VB	NNP	PREP	NNP	.
Simplifying the tags	V	N	to	N	.
Extracted pattern	(give, V n to n, 0), (give, V n, 0)				

Table 1: Process of grammar pattern extraction. Two grammar patterns are extracted for the given phrase. The three columns of the final output indicate the headword, the grammar pattern, and the location of the headword in the original sentence

Headword	**Edit Rule**	**Freq.**
graduate	V at n → V from n	124
graduate	V n → V from n	482
graduate	V in n → V from n	295
call	V to n → V n	390
call	V for n → V n	101
call	V n → V for n	12
ask	V n → V for n	533
ask	V for n → V n	138
ask	V to n → V n	303
talk	V with n → V to n	256
talk	V to n → V on n	217
talk	V n → V to n	385
talk	V n → V on n	156
talk	V n → V about n	119
introduce	V n for n → V n to n	45
introduce	V n n → V n to n	63
discuss	V about n → V n	144
discuss	V with n → V n	17
thank	V for n → V n	118
thank	V for n → V n for n	89
thank	V quote n → V n	61

Table 2: Example of our result.

guistic Data Consortium (LDC). Learners are divided into three categories due to their English proficiency level. We randomly select ten essays for each category and let our English expert correct these essays with the help of our correction rules, aiming to mimic the situation of a teacher correcting students' essays.

For low proficiency level, 50% of the essays contain errors that are correctable by our correction rules, 20% for medium, and 40% for high proficiency level respectively. In general, 36% of essays are correctable by using our correction rules. Our result shows that the correction rules extracted by our method assist the process of correcting learners' essays.

Few of the corrected sentences from the high proficiency level category are shown below:

- **Original sentence:** They cannot attend *to* the social events or community services which generally take place in the cities.
- **Corrected sentence:** They cannot attend the social events or community services which generally take place in the cities.

The sentence above applies the rule *"attend, V to N → V N"*, while the following sentence applies the rule *"spend, V N for N → V N on N"*.

- **Original sentence:** Moreover, students now have to spend too much time *for* preparing for this hard education in order to be successful.
- **Corrected sentence:** Moreover, students now have to spend too much time *on* preparing for this hard education in order to be successful.

5 Conclusion

Our method could be easily adjusted to adapt on different reference data. By combining various annotated learner corpus, the quantity and quality of correction rules could be larger and higher. The result could be used to assist ESL students in avoiding grammatical errors, and aid teachers in correcting students' essays. Additionally, it could be used in the compilation of collocation error dictionaries and the construction of grammar error correction systems. Our pattern extraction algorithm could be used independently for corpus researches (Lin and Shen, 2021). From linguistic aspect of view, the choice of preposition usually depends on the semantic category of the following noun. Future works could be conducted

to investigate the relationship between prepositions and the semantic category of adjacent nouns and verbs.

References

Emily Bender, Dan Flickinger, Stephan Oepen, Annemarie Walsh, and Timothy Baldwin. 2004. Arboretum: Using a precision grammar for grammar checking in call.

Chris Brockett, William Dolan, and Michael Gamon. 2006. Correcting esl errors using phrasal SMT techniques.

Jhih-Jie Chen, Jim Chang, Yang Ching-Yu, Mei-Hua Chen, and Jason S. Chang. 2017. Extracting formulaic expressions and grammar and edit patterns to assist academic writing. In *Proceedings of EUROPHRAS 2017, Computational and Corpus-based Phraseology: Recent advances and interdisciplinary approaches*, pages 95–103, London, UK. Tradulex.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic. Association for Computational Linguistics.

Jens Eeg-Olofsson and Ola Knutsson. 2003. Automatic grammar checking for second language learners —the use of prepositions.

J. Geertzen, T. Alexopoulou, and A. Korhonen. 2013. Automatic linguistic annotation oflarge scale l2 databases: The ef-cambridge open language database(efcamdat). In *Selected Proceedings of the 31st Second Language Research Forum (SLRF)*, Cascadilla Press, MA.

Anta Huang, Tsung-Ting Kuo, Ying-Chun Lai, and Shou-de Lin. 2010. Identifying correction rules for auto editing. In *ROCLING 2010 Poster Papers*, pages 251–265, Nantou, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Y. Huang, A. Murakami, T. Alexopoulou, and A. Korhonen. 2018. Dependency parsing of learner english. international journal of corpus linguistics. In *International Journal of Corpus Linguistics*.

Fu-Ying Lin and Kuan-Yu Shen. 2021. Features of the spoken academic english (of MOOCs): take the grammar patterns of verbs as an example. In *Corpus Linguistics International Conference 2021*.

Guihua Sun, Xiaohua Liu, Gao Cong, Ming Zhou, Zhongyang Xiong, FGFH EWR, and Chin-Yew Lin. 2007. Detecting erroneous sentences using automatically mined sequential patterns. In *Annual Meeting-Association for Computational Linguistics*, volume 45.

ND Turton and JB Heaton. 1996. *Longman Dictionary of Common Errors (New Ed)*. Longman, England.