Exploring Human-judged and Automatically-induced Correction Difficulty for Grammatical Error Correction

Anonymous ACL submission

Abstract

While grammatical error correction (GEC) has improved in its correction performance, one of the key challenges in GEC research still remains in evaluation. Specifically, all errors are equally treated in the conventional performance measures despite the fact that some errors are more difficult to correct than others. Ideally, difficult errors should be regarded to be more important than easy ones in evaluation. This leads to the following ultimate research question — Can even human experts estimate correction difficulty well? In this paper, we explore questions about correction difficulty centering on this research question. For this purpose, we first introduce a method for estimating agreement rates in correction difficulty judgements based on pairwise comparison. With the annotation of 2,025 instances using this method, we show that human experts exhibit a moderate agreement rate of 66.39% (Cohen's-κ: 0.42) in judging correction difficulty. We also show that the agreement between this human-based difficulty and an automatically induced difficulty is comparable (64.50% and $\kappa = 0.35$ on average). We further look into the annotation results to reveal the insights of the human-judged and machine-judged correction difficulties, reporting on following three findings: (i) where the human-judged and machine-judged difficulties are strong and weak; (ii) based on (i), correction difficulty can be GEC-algorithm- and training-corpus-dependent; (iii) human-judged and machine-judged correction difficulties complement each other.

1 Introduction

Recent progress in grammatical error correction (GEC) makes it possible to correct a wide variety of grammatical errors as can be seen in the work by, for example, Omelianchuk et al. (2020); Rothe et al. (2021); Stahlberg and Kumar (2021), to name a few. Such errors range from easy ones to correct (e.g., *It is* **more easy* → *easier*.) to more difficult ones (e.g., *It is difficult for* **the* → ϕ *students.*).

One of the key challenges in research in GEC remains in evaluation. Namely, all errors are equally treated in the conventional performance measures such as $F_{0.5}$ and GLEU (Napoles et al., 2015) despite the fact that some errors are more difficult to correct than others. Ideally, difficult errors should be regarded to be more important than easy ones in evaluation. Nevertheless, there is almost no method satisfying this requirement. An exception is a difficulty-weighted performance measure for GEC proposed by Gotou et al. (2020). Their method automatically estimates correction difficulty based on the success rate of the correction, that is, the proportion of GEC systems that successfully correct the target error to the entire system set in question. What is missing in their method is whether or not the correction difficulty defined in their work reflects well the human judgement of correction difficulty (hereafter, the former and latter will be referred to as *machine-judged correction difficulty* and *human-judged correction difficulty*, respectively).

This leads to the following ultimate research question — **Can even human experts estimate correction difficulty well in the first place?** Cases can easily be found where it is not straightforward at all to determine their correction difficulty by a psychometric scale such as the Likert scale (e.g., *very difficult*, *difficult*, *standard*, *easy*, or *very easy*). For instance, it is not trivial at all to determine the rating of the above first example: *It is* **more easy* → *easier*.; it could be rated as *standard* or *very easy*. Similarly, the same argument applies to the second example: *It is difficult for* **the* → ϕ *students.*, which can be any of *standard* to *very difficult*.

The above ultimate research question brings out further questions related to correction difficulty.

For example, if human experts estimate correction difficulty well, is the machine-judged correction difficulty associated with the human-judged correction difficulty well? Are there any differences between them? If yes, where and how?

In this paper, we explore these questions about correction difficulty centering on the above ultimate research question. To overcome the problem of judging correction difficulty manually, we introduce a method for estimating correction difficulty based on pairwise comparison. This method facilitates the judgement by exploiting the machine-judged correction difficulty. We apply this method to 2,025 error pairs sampled from the CoNLL-2014 shared task test set (Ng et al., 2014) to estimate the agreement rate of the correction difficulty judgement by human experts. We also investigate how well the human-judged correction difficulty agrees with the machine-judged correction difficulty, which in turn reveals advantages and disadvantages of the two.

The contributions of this work are summarized as follows. First, with the proposed method, we show that human experts can indeed estimate correction difficulty to some extent. To be precise, we show that two human experts achieve an agreement rate of 66% (kappa 0.42); the agreement rate rises up to 96% as difference in correction difficulty increases. We then show that the human-machine agreement rate slightly lower, but comparable to the human-human agreement rate (64%; kappa 0.35 on average). We further investigate the judgement results to reveal the following three findings: (i) where the human-judged correction and machine-judged correction difficulties are strong and weak; (ii) based on (i), correction difficulty can be GEC-algorithm- and training-corpus-dependent; (iii) human-judged and machine-judged correction difficulties complement each other.

2 Related Work

With the advent of the deep neural network techniques, GEC has dramatically improved in correction performance. Examples include the work by Omelianchuk et al. (2020); Rothe et al. (2021); Stahlberg and Kumar (2021), to name a few.

In GEC, $F_{0.5}$ (based on recall and precision) and GLEU are widely used as performance measures. In addition, evaluation tools including the Max-Match (M^2) scorer (Dahlmeier and Ng, 2012) and ERRANT (Bryant et al., 2017; Felice et al., 2016) are available to the public. These measures and tools have contributed to progress in GEC. None of these conventional measures nor tools, however, do not consider correction difficulty.

The measure proposed by Gotou et al. (2020) takes correction difficulty into account. Their measure is based on the success rate of GEC. For this purpose, system outputs are first aligned to the corresponding reference sentences. Then, the number of successful corrections are counted to calculate the success rate. Finally, each error is weighted according to its success rate; basically, the lower the success rate is, the more difficult the error is considered to be. Gotou et al. (2020) demonstrate that the weights based on the success rate can be interpreted as correction difficulty. We exploit this correction difficulty to facilitate pairwise comparison.

Numerous corpora are available for GEC evaluation. These include the CoNLL-2013 (Ng et al., 2013) and CoNLL-2014 (Ng et al., 2014) datasets, Cambridge ESOL First Certificate in English (FCE) (Yannakoudakis et al., 2011), JHU FLuency-Extended GUG Corpus (JFLEG) (Napoles et al., 2017), Konan-JIEM Learner Corpus (KJ) (Nagata et al., 2011). These corpora differ in many aspects: proficiency levels and mother tongues of the writers, essay topics, and error rates.

3 Method and Conditions

3.1 Method

The main purpose of this paper is to answer the research question: **can even human experts estimate correction difficulty well?** As described in Sect. 1, the major obstacle to this goal is that it is not straightforward at all to determine correction difficulty by a psychometric scale such as the Likert scale.

To overcome this problem, we propose a method based on pairwise comparison. This is because we found in a pre-experiment that it was much easier to determine which was more difficult given a pair of error instances than to rate individual errors on a Likert scale. For example, one can tell that the first example in Sect. 1 is easier to correct than the second example. For this reason, we adopt pairwise comparison as our basis.

The procedure of the method based on pairwise comparison is summarized in the following three

Error 1	Result	Error 2
It is *more easy → easier to...	<	It is difficult for *the → ϕ ...
*A → The doctor said ...	>	The number *coresponds → corresponds ...
It can be *improve → improved ...	=	It can be *explain → explained ...
*Some how → Somehow I must find ...	?	*May be → Maybe I go to ...

Table 1: Examples of pairwise comparison.

steps:

Step 1 Create pairs to be judged

Step 2 Judge correction difficulty by pairwise comparison

Step 3 Estimate agreement rate

In Step 1, we create pairs of two errors. As a simple way of creating pairs, we can randomly sample two sentences containing errors from learner corpora annotated with grammatical errors such as the CoNLL-2014 shared task test set (Ng et al., 2014). Since a sentence can contain multiple errors, we highlight the target errors to be compared as shown in the example sentences in Table 1.

Although this simple way enables us to create pairs for evaluation, it is far from efficient. The resulting pairs should vary in terms of the difference in correction difficulty between each pair. In other words, it is better to include various correction difficulties in the pairs, for example, a pair consisting of *very difficult* and *very easy*, *very difficult* and *easy*, and so on. Random sampling does not satisfy this requirement unless the distribution of correction difficulty pairs is not uniform.

To improve the efficiency, we exploit the method for automatically estimating correction difficulty introduced by Gotou et al. (2020), that is the machine-judged correction difficulty. With this method, we can assign a (tentative) correction difficulty rating to each error in a learner corpus. We can then sample an error in one rating category and another in another rating category and make them a pair instead of randomly sampling two errors from the entire set. Doing so, we control the variety of the difference in correction difficulty in the pairs for evaluation.

One might argue that we cannot use the automated method because it has not yet been proven to correlate to the human-judged correction difficulty. However, even in the worst case (i.e., the automated method does not estimate correction difficulty well at all, and thus randomly outputs one of the correction difficult ratings), it would only result in the situation where all errors are sampled randomly as in the simple way described above. Besides, we can tell if this is the case or not by looking at the judgement results. In this worst case, the human-machine agreement should greatly deviate from the human-human agreement.

In Step 2, the obtained pairs are displayed to human experts, who judge correction difficulty by either *the first error is more difficult to correct*, *the second is more difficult*, *equally difficult*, or *cannot judge*, which are denoted by $>$, $<$, $=$, and ? in Table 1.

Before we actually judged correction difficulty in the above manner, we had conducted a trial session to make judgement criteria. They are summarized as follows:

C1 Amount of context

C2 Lexicality

C3 Multiple errors

C1 refers to the amount of context required to correct the error in question; the wider the context is, the more difficult the error is considered to be. For example, although the following two error instances fall into the same error category (i.e., subject-verb agreement), the former requires a wider context (five words to the subject *students*), and thus it is considered to be more difficult to correct: *The students in the new class *likes → like their teacher.* vs. *The students *likes → like reading.* The amount of context is calculated based on the number of words (to the clue). If the clue is beyond the sentence boundary, it is considered to be more difficult than those inside the sentence in question. Likewise, if the clue is extra-textual, the amount of context required is regarded as infinite (and thus more difficult than the other two cases).

C2 concerns the lexicality of the correction. Errors involving lexical choice tend to be more difficult. For example, the error *Can you *teach → tell*

me the way to the station. involves lexical choice and thus is expected to be more difficult than the error *He *tell → tells me what to do.* Note that C2 is often associated with C1; errors involving lexical choice require a wider context. For example, the first error requires almost the whole sentence to correct *teach* to *tell* whereas the second can be corrected by just looking at *He tell*.

C3 is used when other errors appear around the error in question. To be precise, it is considered to be more difficult if correcting the error in question is influenced by other errors. For example, the error *A students *likes → like it.* would be much easier to correct without the other error as in *Students *likes → like it*.

All these criteria are of course not the gold standard rules and have room for interpretation. Multiple criteria may sometimes apply to the same error simultaneously, in which case one has to decide which one is superior. We let human experts decide the final choice based on these criteria.

Finally, in Step 3, we estimate the agreement rate based the obtained judgements. We simply define it as the number of pairs whose judgement results are agreed by two human experts (excluding *cannot-judge* cases) divided by the total number of pairs. We also use Cohen's-κ (Cohen, 1960) as another estimate.

It should be emphasized here that as well as the agreement rate between human experts, we can estimate human-machine agreement rates in the same manner. In the above case, the agreement is determined based on the human judgements. In contrast, in this case, it is determined whether the machine judgement agrees with its human counterpart. Recall that pairs are created according to the machine-judged correction difficulty (and thus it tells which is difficult).

3.2 Conditions

We use the widely used CoNLL-2014 shared task test set (Ng et al., 2014) as our base learner corpus. We only use the first annotation, which contains 2,379 errors.

To create error pairs (i.e., Step 1 in Sect. 3.1), we need to automatically estimate correction difficulty of the above errors (i.e., to implement the Gotou et al. (2020)'s method). We in turn need to obtain error correction results for the corpus, which are used to implement their method. We choose the same eight systems used in their work: specifically, a phrase-based statistical machine translation-based system (Junczys-Dowmunt and Grundkiewicz, 2016), three deep neural network-based systems (Junczys-Dowmunt et al., 2018; Ge et al., 2018; Kiyono et al., 2019), and four baseline systems, which are introduced by Mita et al. (2019), based on statistical machine translation-based or deep neural networks. Note that the use of the eight systems provides nine levels of correction difficulty. Table 2 shows their distribution. In Table 2, higher values denote more difficult errors.

We use a subset of all pairs obtained by using the method described in Sect. 3.1. Specifically, we randomly choose 50 pairs for each combination of different difficulty levels. The nine levels of correction difficulty make 36 combinations of them and thus it makes 1,800 pairs of error instances. We also include pairs whose difficulty levels are the same. Simply, we randomly choose 25 pairs from each difficulty level, which amounts to 225 pairs. Accordingly, we use 2,025 error pairs in total. The resulting pair contain 1,412 unique errors; note that the same errors are inevitably used multiple times because the number of errors differ depending on the difficulty level.

The second and third authors conduct the difficulty judgement. They have been engaged on GEC research for more than 20 years and five years, respectively and have developed a number of GEC systems. They independently conduct pairwise comparison (i.e., Step 2 in Sect. 3.1) for the resulting pairs. After the first round, they recheck the results (again, independently) to reduce annotation mistakes.

Difficulty level	Frequency
0	110
1	80
2	104
3	107
4	109
5	111
6	154
7	208
8	1,396
Total	2,379

Table 2: Distribution of machine-judged correction difficulty.

Evaluator pair	Agreement rate (%)	Cohen's-κ
H_1 - H_2	65.38	0.41
H_1 - M	59.36	0.32
H_2 - M	58.37	0.28
H_1 - H_2	66.39	0.42
H_1 - M	64.72	0.37
H_2 - M	64.28	0.33

Table 3: Simple agreement rate and Cohen's-κ in pairwise comparison. Upper block: results for all 2,025 pairs; Lower block: results excluding equivalent pairs; H_1: human expert 1; H_2: human expert 2; M: machine-judged correction difficulty.

	Judgement			
Evaluator	<	=	>	?
H_1	922	256	846	1
H_2	938	167	920	0

Table 4: Distribution of difficulty judgements.

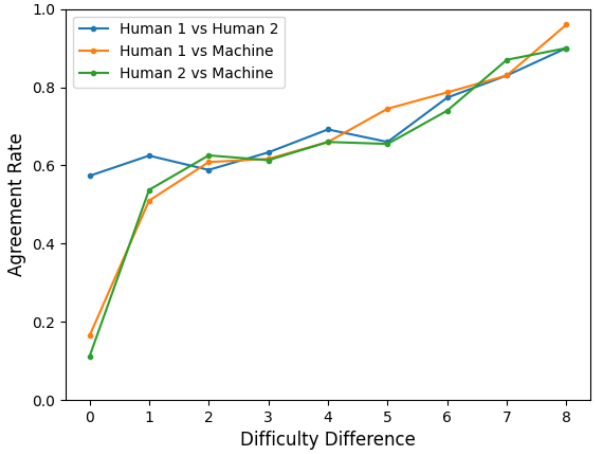

Figure 1: Relationship between difference in correction difficulty and simple agreement rate.

4 Result

Table 3 shows the simple agreement rates and Cohen's-κ; Table 4 shows the distribution of the judgments (pairwise comparison). In both tables, H_1, H_2, and M denote the first, second human experts, and the machine-judged correction difficulty, respectively. The upper and lower blocks in Table 3 correspond to the results excluding and including the error pairs whose difficulty is equivalent, respectively.

According to Table 3, the human judgements exhibit moderate agreement. This agreement rate of 65.38% is significantly higher than the majority baseline (44.44%) (two-proportion z-test; $p < 0.01$). At the same time, Table 3 suggests that there are cases on which even human experts disagree, which will be discussed in detail in Sect. 5.1.

Table 3 also shows that the human-machine agreements are slightly lower than, but comparable to the human-human agreements especially in the results excluding the 225 difficulty-equivalent pairs. This implies that human experts can judge correction difficulty with a finer grade. Namely, they can tell the difference even in cases where the automated method cannot.

A closer look at the results reveal insights of the human and machine judgements. Figure 1 shows the relationship between the difference in correction difficulty and the simple agreement rate. The horizontal and vertical axes correspond to the difference in correction difficulty and the simple agreement rate, respectively. The overall trend is that the agreement rate goes higher in all pairwise comparison as the difference in correction difficulty increases. In particular, for the error pairs with the largest difficulty difference (i.e., difference 8), the corresponding agreement rates exceed 90%. The plots at the difference 0 in Figure 1 also confirm the previous argument that human experts can judge correction difficulty with a finer grade.

The results are summarized as follows. The obtained simple agreement rates and Cohen's-κ gives (at least to some extent) *yes* to the research question – can even human experts estimate correction difficulty well in the first place? The machine-judged correction difficulty is expected to be comparable to the human-judged correction difficulty especially when the difference in correction difficulty are relatively large.

5 Discussion

5.1 Sources of Disagreements

In Sect. 4, we have seen that the human judgements exhibit moderate agreement. We also have

seen that the human-machine agreements are comparable to the human-human agreements. Unfortunately, however, both cases do not achieve perfect agreement.

We now turn to the question where the disagreements come from and why. To discuss this point, let us first look at the simple agreement rates by error types, which are shown in Table 5. Here, the error types are those automatically obtained by using ERRANT (Bryant et al., 2017); see their paper for the error type definition. Note that the 225 error pairs whose correction difficulty is equivalent are excluded from Table 5. Also note that OTHER (other type) and error types whose occurrences are less than ten are excluded.

To our surprise, both human-human and human-machine agreement rates exhibit similar values in most of the error types. This may reflect the fact that ambiguous, subtle cases that are difficult to judge according to the machine-judged correction difficulty are also difficult for the human experts, and vice versa. At both extremes, for example, the first two error types exhibit a very high agreement rate in all three combinations; the first (ADJ) tends to be judged as more difficult to correct whereas the second is the opposite case. Most of the other cases such as VERB:TENSE (tense error) and DET (determiner errors) also exhibit similar values across the three, but their values are much lower, suggesting that there are cases that are difficult to judge for both human experts and the automated method. More generally, the judgement of correction difficulty can be highly difficult for even human experts in some cases.

Interestingly, there are error types whose agreement rates considerably differ in the three. The most typical case is found in ORTH, specifically, errors concerning white spaces (e.g., **thefamily* → *the family* and **some how* → *somehow*). Although its agreement rate between one of the human experts and the machine-judged correction difficulty is as high as 70%, the human-human agreement reaches only 30%. The first expert tends to rate this type of error as easier to correct because the rule for correcting this type of error is rather simple and clear. The other expert, however, favors the opposite considering that modern systems based on neural networks or even statistical machine translations do not normally take account of such a rule. This implies that correction difficulty can be (at least partially) GEC algorithm-dependent, which should be one of the factors that makes human judgement difficult.

	Simple Agreement Rate		
Error Type	H_1 - M	H_2 - M	H_1 - H_2
ADJ	0.92	0.85	0.92
VERB:INFL	0.90	1.00	0.90
WO	0.88	0.75	0.69
CONJ	0.87	0.73	0.73
PUNCT	0.75	0.80	0.81
NOUN	0.73	0.72	0.58
ADV	0.72	0.72	0.72
VERB	0.72	0.68	0.70
PART	0.70	0.63	0.59
ADJ:FORM	0.69	0.46	0.54
VERB:FORM	0.69	0.65	0.68
PRON	0.67	0.72	0.72
VERB:TENSE	0.67	0.70	0.69
PREP	0.66	0.64	0.70
VERB:SVA	0.64	0.62	0.65
NOUN:INFL	0.64	0.64	0.50
MORPH	0.63	0.58	0.60
DET	0.62	0.61	0.62
NOUN:NUM	0.59	0.58	0.61
NOUN:POSS	0.56	0.56	0.69
SPELL	0.54	0.61	0.77
ORTH	0.45	0.70	0.30

Table 5: Simple agreement rates for each error type (excl. 225 equivalent pairs).

The differences in the simple agreement rates is also large in SPELL (spelling errors). The agreement rate between the two human experts is relatively high compared to the human-machine agreement. They frequently judge spelling errors, which appear 211 times in the data, to be easier than their counterpart (the error for comparison). This is only natural if we consider the criterion C1 that the wider context an error requires to be corrected, the more difficult it is considered to be. It should be emphasized that most spelling errors can be corrected without any context (by the word itself). An actual example is: $\cdots$ *it is a good practice not to* **intesively* → *intensively* *use social media all the time.* Humans can almost immediately correct the error by just looking at the target word *intesively*. In the pairwise comparison experiment, this error instance is paired with the following error in subject-verb agreement: $\cdots$ *with the function of social media*

sites that *connects → connect *the people* , · · · . In order to correct this error, one needs to recognize that the target word is a verb and that its subject is *social media sites* (and not *function* nor *that*). The former case is much simpler as a correction procedure than the latter. Nevertheless, none of the eight GEC systems successfully corrects the former (spelling error) while half of them succeed in the latter (subject-verb agreement error). Accordingly, the former is judged to be easier in the machine-judged correction difficulty, which disagrees with the human judgements. This is the major source of the human-machine disagreement. For instance, 19 spelling errors are successfully corrected by only two or less of the eight GEC systems and are judged to be more difficult than their counterpart. Out of 19, 17 are judged as the opposite by both human experts.

These observations about spelling errors imply that correction difficulty can be training-corpus-dependent as well as GEC algorithm-dependent. Even spelling errors that are easy to correct for humans can be very difficult for corpus-based GEC systems if they never appear in the training corpus. Admittedly, recent deep neural network-based systems are based on subwords and are influenced less by unseen spellings. That said, it would be very hard to correct unseen spelling errors with standard deep neural network-based GEC systems. This exemplifies that correction difficulty can be training-corpus-dependent. The same argument can partly apply to the above orthographic errors. Of course, unseen spelling and orthographic errors can mostly be corrected with neural network-based GEC systems if they are equipped with specialized functions. This reflects a GEC algorithm-dependent aspect of correction difficulty.

So far, we have observed that judgements of correction difficulty can be very difficult for both human experts and the automated method in some cases. This is partly ascribed to the training-corpus-dependent and the GEC algorithm-dependent aspects of correction difficult, which are sources of disagreement in judgements.

5.2 Advantages and disadvantages of human-judged and machine-judged correction difficulties

The results shown in Sect. 4 suggest that the human-judged correction difficulty exhibits a slightly higher agreement rate and is finer-graded than the machine-judged correction difficulty. At the same time, it is highly costly and time-consuming to manually annotate grammatical errors with their correction difficulty even with the pairwise comparison adopted in this paper. It often requires human experts to conduct pairwise comparison accurately. For this reason, the human-judged correction difficulty is more suitable for deep analysis of grammatical errors in terms of correction difficulty.

In contrast, the machine-judged correction difficulty has an advantage over the human-judged correction difficulty in terms of cost and time. In other words, it enables us to assign correction difficulty ratings to a large number of error instances with a much shorter time, which is preferable or even necessary in certain situations such as evaluation in GEC shared tasks where a number of systems are involved and/or where the test set is large.

In Subsect. 5.1, we have observed that in some cases, the machine-judged correction difficulty reveals insights of GEC systems that even human experts are not aware of. Specifically, it has revealed that there exist rather simple errors (simple in terms of error correction) that even sophisticated, state-of-the-art GEC systems cannot correct, suggesting that it will be useful to strengthen manual analysis of correction difficulty.

This nice property also suits the machine-judged correction difficulty evaluation in GEC shared tasks. The reasons for this are (i) the performance measure based on the machine-judged correction difficulty gives higher weights to errors that are not corrected by other systems in evaluation as Gotou et al. (2020) (and also the above discussion) demonstrate; (ii) this means that researchers in the domain of GEC have to tackle such challenging errors to achieve a better performance; (iii) this in turn brings a diversity of GEC systems.

The reader might be wondering how many GEC systems are required to achieve a stable difficulty judgment in the machine-judged correction difficulty; we used eight GEC systems in this work, which is not so cost- and time- effective although

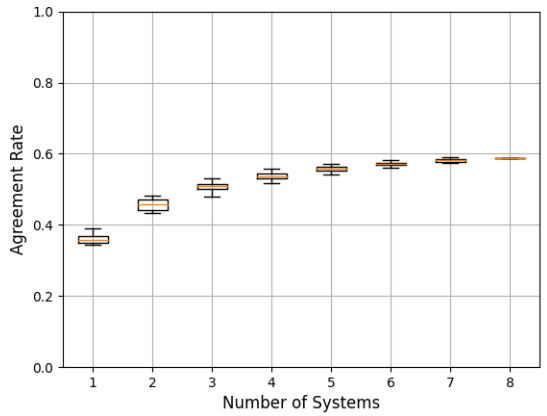

Figure 2: Relationship between the number of GEC systems used and human-machine agreement rates.

we can use already-implemented systems repeatedly in practice.

Figure 2 partly answers this question. The horizontal and vertical axes of the figure indicate the number of GEC systems used to calculate the machine-judged correction difficulty and the average of corresponding simple agreement rates. The average is taken over the possible combination of used systems (28 combinations when two systems are used, for example) and the two experts.

As expected, the average human-machine agreement rate improves as the number of GEC systems used increases. More importantly, the figure shows a saturation point comes at around five or six systems. In other words, the more system used, the better. At the same time, we can obtain comparable results with five or six systems. Of course, this only applies to this dataset (i.e., the CoNLL-2014 shared task test set) with the eight GEC systems. The tendency might greatly change depending on datasets and GEC systems. It will be interesting to explore the relationship in detail.

6 Conclusions

In this paper, we have explored research questions about correction difficulty in GEC. To answer the questions, we first introduced a method for estimating correction difficulty efficiently. With the annotation of 2,025 instances, we showed that human experts exhibit moderate agreement rate of 66.39% (Cohen's-κ: 0.42) while the human-machine agreement rate is comparable (64.50% and $\kappa = 0.35$ on average). We further looked into the annotation results to reveal insights of human-judged and machine-judged correction difficulties. Specifically, we reported on the following three findings: (i) where the human-judged and machine-judged difficulties are strong and weak; (ii) based on (i), correction difficulty can be GEC-algorithm- and training-corpus-dependent; (iii) human-judged and machine-judged correction difficulties complement each other.

References

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Tao Ge, Furu Wei, and Ming Zhou. 2018. Reaching human-level performance in automatic grammatical error correction: An empirical study. *arXiv preprint arXiv:1807.01270*.

Takumi Gotou, Ryo Nagata, Masato Mita, and Kazuaki Hanawa. 2020. Taking the correction difficulty into account in grammatical error correction evaluation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2085–2095, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Austin, Texas. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a

low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China. Association for Computational Linguistics.

Masato Mita, Tomoya Mizumoto, Masahiro Kaneko, Ryo Nagata, and Kentaro Inui. 2019. Cross-corpora evaluation and analysis of grammatical error correction models — is single-corpus evaluation enough? In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1309–1314, Minneapolis, Minnesota. Association for Computational Linguistics.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1219, Portland, Oregon, USA. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.