Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5565–5572
Marseille, 20-25 June 2022

Construction of a Quality Estimation Dataset for Automatic Evaluation of Japanese Grammatical Error Correction

Daisuke Suzuki[1], Yujin Takahashi[1], Ikumi Yamashita[1], Taichi Aida[1], Tosho Hirasawa[1], Michitaka Nakatsuji[1], Masato Mita[2,1], Mamoru Komachi[1]
Tokyo Metropolitan University[1]
RIKEN[2]
{suzuki-daisuke3@ed., takahashi-yujin@ed., yamashita-ikumi@ed., aida-taichi@ed., hirasawa-tosho@ed., nakatsuji-michitaka@ed.,}tmu.ac.jp, masato.mita@riken.jp, komachi@tmu.ac.jp

Abstract

In grammatical error correction (GEC), automatic evaluation of the system outputs is an important factor for the research and development of GEC systems. In this regard, existing studies on automatic evaluation have demonstrated that quality estimation models built from datasets with manual evaluation can achieve high performance in automatic evaluation of English GEC without using reference sentences. However, quality estimation models for the Japanese language have not been studied yet owing to the absence of datasets necessary for constructing such models. Therefore, in this study, we created a quality estimation dataset with manual evaluation to build an automatic evaluation model for Japanese GEC. Moreover, we conducted a meta-evaluation to verify the usefulness of this dataset in building the Japanese quality estimation model.

Keywords: corpus construction, automatic evaluation, quality estimation, Japanese grammatical error correction

1. Introduction

Grammatical error correction (GEC) is the task of correcting different types of errors in text such as those pertaining to spelling, punctuation, grammar, and word-choice. Automatic evaluation is considered as an important factor to enable the continuous integration and deployment of GEC systems.

There are two types of automatic evaluation systems for GEC: reference-based and reference-less. Reference-based systems have been used extensively by the GEC community, whereas reference-less systems have gained considerable attention only recently. Hereafter, we refer reference-less systems as *quality estimation models*.

One of the problems for reference-based methods is the diversity of references. In a study by (Bryant and Ng, 2015), a dataset with multiple corrections was created by 10 annotators based on the CoNLL-2014 test set (Ng et al., 2014). The authors observed that the number of error types corrected by each annotator varied significantly. This means that there can be a variety of corrected versions for a grammatically incorrect sentence. However, with a limited number of reference sentences, it is difficult to cover a large variety of corrected sentences in GEC. Therefore, reference-based automatic evaluation methods may assign unreasonably low scores to valid system-outputs when they are not observed amongst the references.

By contrast, quality estimation models do not present this problem because they estimate the quality of the system's output without requiring gold-standard references. Asano et al. (2017) and Yoshimura et al. (2020) built a quality estimation model using a dataset with manual evaluations of system-outputs,[1] and achieved a higher correlation between manually evaluated values and automatically assigned scores than that by reference-based methods.

However, there is no dedicated quality estimation dataset for the Japanese language, although Japanese is popular amongst learners. As a result, no reference-less automatic evaluation method has been proposed for Japanese GEC. The NAIST Lang-8 Learner Corpora (Mizumoto et al., 2011) is one of the largest corpora of texts written by language learners. Notably, out of the 580,549 essays in this corpus, 192,673 (which is the second-largest number after English) are written in Japanese. This indicates that a study focusing on quality estimation models for Japanese GEC can make a large impact on the GEC community.

Therefore, in this study, we created a quality estimation dataset with manual evaluation to build an automatic evaluation model for Japanese GEC. The dataset consisted of three components: source text, corrected texts, and manual-evaluation scores. The source text is derived from the Lang-8 corpus, and the corrected texts consist of the outputs of four diverse GEC systems. We built a quality estimation model by fine-tuning a pre-trained sentence encoder (namely BERT (Devlin et al., 2019)) on the created dataset, and calculated the correlation with the manual-evaluation values. Additionally, we calculated the correlations with manual evaluations for reference-based automatic evaluation methods and compared them against the correlations of the proposed quality estimation model to meta-

[1] Asano et al. (2017) have built several quality estimation models, and one of them uses a dataset comprising manual-evaluation scores.

evaluate the quality estimation model built using this dataset.

The main contributions of this study are as follows.

- To build a quality estimation model for Japanese GEC, we created a dataset comprising outputs from multiple GEC systems that were annotated via human evaluation.
- We demonstrated that the quality estimation model for Japanese GEC performs better than reference-based automatic evaluation; this was achieved by building a quality estimation model using the newly created data set and evaluating its performance.

2. Related Work

2.1. Evaluation Method

Reference-based methods. Initially, the outputs of English GEC systems were evaluated by calculating the match rate, recall rate, and F-score for each word (Dale and Kilgarriff, 2011). The success of shared tasks in GEC has led to the large-scale adoption of the maximal matching (Max Match, or M^2) technique by the NLP community; this method calculates the match rate, recall rate, and $F_{0.5}$ score for each phrase. However, despite the prevalence of the M^2 scorer, Felice and Briscoe (2015) proposed the I-measure, which assigned a score between -1 and 1, while other evaluation methods assign a score between 0 to 1; these methods attempt to ensure that the bad corrections will receive low scores while good corrections will receive high scores. Furthermore, Napoles et al. (2015) proposed GLEU, which is a modified version of BLEU (Papineni et al., 2002) for evaluating GEC. BLEU evaluates machine translation system outputs by comparing the translated sentence with the reference sentence, while GLEU evaluates GEC system outputs by comparing three sentences: the source sentence, the system-corrected sentence, and the reference sentences. Amongst the methods that used reference sentences, GLEU achieved the highest correlation with manual evaluation.

As for Japanese GEC, Mizumoto et al. (2011) conducted automatic evaluation using BLEU to compare the grammaticality of the sentences written by a language learner and the reference sentences. More recently, Koyama et al. (2020) conducted automatic evaluation using GLEU with an original evaluation corpus for Japanese GEC constructed during their research.

Reference-less methods. In contrast to reference-based automatic evaluation methods, quality estimation models have been proposed recently. Napoles et al. (2016) evaluated a GEC system based on the number of errors detected by the grammatical error detection system and demonstrated that its performance was at par with that of GLEU. Asano et al. (2017) proposed a method for evaluating corrected sentences in terms of grammaticality, fluency, and meaning-preservation using logistic regression models, RNN language models, and METEOR (Denkowski and Lavie, 2014), respectively. Yoshimura et al. (2020) proposed a method to optimize the three evaluation measures considered in Asano et al. (2017) for manual evaluation. Yoshimura et al. (2020)'s method builds a quality estimation model by finetuning a pre-trained sentence encoder (BERT) on the manual evaluation of each measure and shows better correlation with the manual evaluation than Asano et al. (2017).

However, although reference-based automatic evaluation is commonly used, quality estimation models have not been applied to Japanese GEC (Mizumoto et al., 2011; Koyama et al., 2020). Therefore, to adapt a reference-less automatic evaluation method for Japanese GEC, we created a dataset with manual-evaluation values for building a quality estimation model.

2.2. Dataset with Human Evaluation

We introduce existing datasets with manual evaluation of system-corrected sentences as related work.

The GUG dataset (Heilman et al., 2014) consists of 3,129 sentences sampled randomly from essays by English language learners. The dataset was annotated by two linguistically trained native English speakers. The GUG dataset was created to assess the performance of automatic evaluation methods, and it has also been used as training data for quality estimation models. In this study, the evaluation was conducted using a 5-point Likert scale, following (Heilman et al., 2014).

Grundkiewicz et al. (2015) assigned human ratings to the output of 12 GEC systems that participated in the CoNLL-2014 shared task on English GEC (Ng et al., 2014). In their dataset, the human ratings were annotated relative to the source text based on the ranking of multiple corrections.

In this study, we created a dataset of manual evaluation as the training data for a quality estimation model for Japanese GEC. This novel dataset will be useful for evaluating the performance of automatic evaluation methods for Japanese GEC.

3. Construction of QE Dataset

To construct the dataset, we first created pairs of error-containing and system-corrected sentences. To create sentence pairs, we used four distinct GEC systems (Sec. 3.1) to generate corrected sentences for two corpora of Japanese learners (Sec. 3.2). Then, we manually evaluated the pairs of error-containing and system-corrected sentences (Sec. 3.3).

3.1. Grammatical Error Correction System

We used the NAIST Lang-8 Learner Corpora (Mizumoto et al., 2011) as the training dataset for the four GEC systems. These corpora were created by extracting the essays written by language learners and the corresponding correction logs between 2007 and 2011 from Lang-8, which is a mutual-correction social media platform for language learners. The corpora comprise the sentences by language learners and the corresponding corrections, essay IDs, user IDs, learning-language tags, and native-language tags in the JSON format. The number of essays with the Japanese learning-language tag is 192,623, and the number of sentence pairs of learner-sentences and the corresponding corrections is approximately 1.3 million.

We employed four GEC systems to output four corrected sentences per input sentence to collect manual evaluations of various types of corrected sentences. The following four representative GEC systems were selected based on Yoshimura et al. (2020).

SMT: This model is based on statistical machine translation (SMT), which is a method that learns the translation probability of each word or phrase and the probability of the correct sequence as statistical information. Moses (Koehn et al., 2007) was used as a tool to perform SMT. KenLM (Heafield, 2011) was used to train the language model.

RNN: A sequence-to-sequence transformation model based on a recurrent neural network, which is a neural method that considers information concerning the time series of data. The implementation is based on fairseq (Ott et al., 2019). Experiments were conducted using bi-directional LSTM. The number of word dimensions was set to 512; the batch size was set to 32, and the rest of the implementation settings were the same as those in (Luong et al., 2015).

CNN: A sequence-to-sequence transformation model based on convolutional neural networks that learns by abstracting features of the data. The implementation is based on fairseq. The dimensions of the encoder and decoder were set to 512, and the remainder of the implementation settings were the same as those in (Chollampatt and Ng, 2018).

Transformer: A sequence-to-sequence transformation model consisting only of a mechanism called attention, which represents the attention of words in a sentence. The implementation is based on fairseq. The parameter settings were the same as those of (Vaswani et al., 2017).

3.2. Datasets

To obtain pairs of error-containing and system-corrected sentences, we used the TEC-JL (Koyama et al., 2020) and FLUTEC [2] datasets, respectively.

TEC-JL. TEC-JL includes 139 essays from the NAIST Lang-8 Learner Corpora containing 2,042 sentence pairs between learners' sentences and corrections. In Lang-8, ordinary users make corrections in the absence of annotation guidelines; thus the corrections may contain noise. In contrast, corrections in the TEC-JL dataset were made with minimal editing by native Japanese speakers with an annotation guidelines and discussions concerning annotation to ensure consistent annotation. Thus, unlike the original Lang-8 corpus, the TEC-JL is a relatively reliable dataset for evaluation. In our experiments, we selected this dataset for comparing the performances of our quality estimation model with the original automatic evaluation methods using reference sentences.

Corpus	# of essays	# of sentence pairs
Lang-8 (ja)	192,673	1,296,114
TEC-JL	139	2,042
FLUTEC	169	2,100

Table 1: Number of essays and sentence pairs in the dataset used in the experiment.

FLUTEC. The dataset consists of sentences by Japanese language learners extracted from the NAIST Lang-8 Learner Corpora and their fluency-aware corrections. The dataset consists of development and test data, with 1,050 sentences randomply sampled each. The texts in this dataset were sampled from the same dataset as TEC-JL; however, the sampling was performed manually to avoid overlap. Notably, in the experiment described in Sec. 4, we used only the dataset derived from TEC-JL.

3.3. Annotation

Policies. The four GEC systems corrected the sentences of Japanese learner in the TEC-JL and FLUTEC datasets, thereby yielding a set of pairs of Japanese learner sentences and system-corrected sentences. To create an effective dataset for building a quality estimation model, we conducted annotation based on two policies: 1) Holistic evaluation. 2) Annotation based on pairs of Japanese-learner- and system-corrected sentences.

First, Yoshimura et al. (2020) collected human ratings of system-corrected sentences based on three metrics: grammaticality, fluency, and meaning-preservation. Then, they fine-tuned BERT on each metric to build a quality estimation model. The model trained on grammaticality yielded a higher correlation with human ratings compared to reference-based automatic evaluation methods. Based on these results, to create a quality estimation dataset at a low cost, we collected human ratings using a holistic grammaticality-oriented scale in this study.

Second, in Yoshimura et al. (2020), grammaticality was evaluated by examining only the system-corrected sentence using a simple five-step evaluation rule. By contrast, to consider meaning-preservation on a single evaluation scale, pairs of original and system-corrected sentences were evaluated herein, and the rules were designed based on the evaluations of both before and after the correction.

[2] https://github.com/kiyama-hajime/FLUTEC

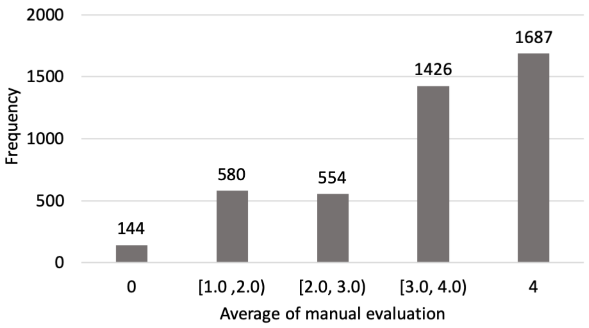

Figure 1: Histogram of manual evaluation scores for sentence pairs generated from TEC-JL.

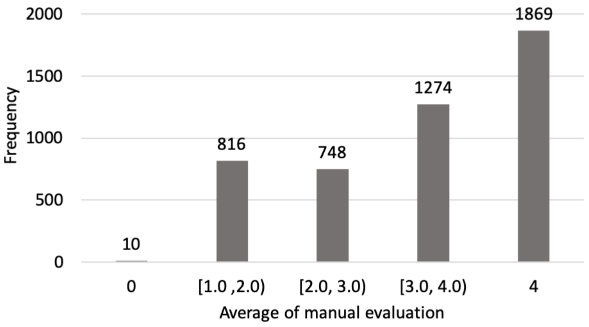

Figure 2: Histogram of manual evaluation scores for sentence pairs generated from FLUTEC.

Process. Four GEC systems were used to generate system-corrected sentences for the 2,042 Japanese learner-written sentences included in TEC-JL. After excluding duplicated sentence pairs, we obtained 4,391 unique sentence pairs. We asked three native Japanese-speaking university students to evaluate the 4,391 sentence pairs. For improving agreement among the annotators, the evaluations up to 2,000 sentences were discussed. The cases where scores differed by more than two were discussed among the annotators, and the guidelines were updated to supplement the rules for cases where evaluations were likely to differ.

Guidelines. We used a 5-point Likert scale scheme to annotate grammaticality-based evaluation scores. The evaluation scores were determined according to the rules described in Table 2. [3]

In the experiment, the averaged ratings of the three annotators were used as the final manual evaluation. As an exception, to reflect the annotator's evaluation accurately, cases where zero was included in the three evaluations were handled as follows: cases when one of the three annotators rated the case as 0, the average of the other two annotators' ratings was taken; when two or more annotators rated the case as 0, no average was taken but the rating was set to 0.

3.4. Analysis

To measure the agreement rate between annotators, we used the kappa coefficient. The value was found to be 0.49, indicating moderate agreement (Landis and Koch, 1977). Figures 1 and 2 show histograms of manual evaluation scores for sentence pairs created from TEC-JL and FLUTEC.[4] For both datasets, the overall number of ratings ranges between 3 and 4. In the graph for TEC-JL, the limited number of 0 ratings is probably because low noise cases were selected by random sampling; for FLUTEC, the number of 0 ratings is limited because low noise data was manually selected during sampling.

Table 3 shows actual annotation examples. In the above annotation example, a wrong sentence was only partially corrected. The annotator categorized this failure as a minor error, and rated the correction 2 or 3 at the discretion of the annotator. In the annotation example below, all annotators gave a rating of 4 because the valid and sufficient corrections were made.

4. QE Experiments

To evaluate the quality estimation performance of the Japanese GEC, we measured the sentence-level correlation coefficients between the human evaluation scores of the reference-based and reference-less evaluations of the output of the GEC system.

4.1. Settings

To build a quality estimation model, we used the same method as Yoshimura et al. (2020) to fine-tune BERT. The input of BERT is the language learner's written and system-corrected sentences, and the output is the evaluation score. Because Yoshimura et al. (2020) assessed grammaticality, fluency, and meaning- preservation separately, only the system-corrected sentences were used as the input to the quality estimation model for grammaticality. However, in this study, because we also assessed meaning-preservation and grammaticality simultaneously, both the learner's sentences and system-corrected sentences were used as the input for BERT. We changed the output layer of the BERT to a regression model to output the ratings.

As for the proposed method, we performed a 10-fold cross-validation using the dataset created in Section 3.3. We divided the dataset into 10 parts so that the ratio of training, development, and testing was 8:1:1, and we fine-tuned BERT with the training data. Using

[3] In addition, we did not consider the following usages. First, the conversion from present tense to past tense is not treated as an error. Second, commas (“，”) and periods (“．”) were treated the same as reading (“、”) and punctuation (“。”) marks. Third, we ignore emoticons and symbols regardless of their position.

[4] Because of the way averages are taken, there is no rating greater than 0 and less than 1.

Score	Description
4	S2 is grammatically correct and all significant and minor errors in S1 have been corrected.
3	All significant errors in S1 have been corrected. Some minor errors are acceptable. (e.g., sentences containing not wrong, but unnecessary edits; presence of white-spaces in S2.)
2	Major errors in S1 are corrected in S2, but one or more minor unacceptable minor errors remain uncorrected. However, at least one minor erroneous corrections has been made. (e.g., missing punctuation; sentences ending with a comma. N.B. These errors can have a score of 3 at annotator's discretion.)
1	Both significant and minor errors in S1 remain uncorrected in S2. Severely erroneous corrections have been made (e.g., changes to nouns and verbs that significantly impair the meaning of the original text, cases where the intended meaning of the original text has been modified significantly although the sentences are grammatically correct.)
0	S1 is a sentence that is difficult to correct or impossible to correct. (e.g., sentences where more than half of the text is non-Japanese.)

Table 2: Description of evaluation scores. S1 is the original text and S2 is the corrected text.

Source text	Corrected sentence
この アルバイトを 本当に 楽しみに します。 (this, part-time job, really, look forward, be [future])	この アルバイトを 本当に 楽しみ です。 (this, part-time job, really, look forward, be [present])
Scores of three annotators (avg.): 2, 3, 3 (2.67)	
元々 沖縄へ 行き たい でした。 (originally, Okinawa, go, want [present], be [past])	元々 沖縄へ 行き たかった です。 (originally, Okinawa, go, want [past] , be [present])
Scores of three annotators (avg.): 4, 4, 4 (4.00)	

Table 3: Annotation examples.

Method	Pearson	Spearman
GLEU	0.320	0.362
fine-tuned BERT	**0.580**	**0.413**

Table 4: Results of meta-evaluation.

the development data, we selected BERT hyperparameters with maximum Pearson's correlation coefficient by grid search with maximum sentence lengths of 128 and 256; batch sizes of 8, 16, and 32; learning rates of 2e-5, 3e-5, and 5e-5; and number of epochs between 1 and 10.

For the baseline, we used GLEU as a reference-based method. Because GLEU is evaluated with a value between 0 and 1, the evaluation was multiplied by 4 for comparison. We used two sentences from TEC-JL as reference sentences. For the test data, we calculated the GLEU scores, quality estimation scores, and measured sentence-level correlations via manual evaluations. For meta-evaluation, we used Pearson's produce-moment correlation coefficient and Spearman's rank correlation coefficient between the scores of the automatic and the manual evaluations for each sentence in the test data.

4.2. Results

Table 4 shows the results of sentence-level comparison for each automatic and manual evaluation score using Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient. The results of the experiment show that fine-tuned BERT is more highly correlated with human ratings than GLEU in terms of both correlation coefficients. Figure 3 shows that there is negligible correlation between the GLEU scores and the manual ratings, while the fine-tuned BERT model shows a weak correlation with the manual ratings, and the model tends to perform a proper quality estimation especially for the cases rated 1.

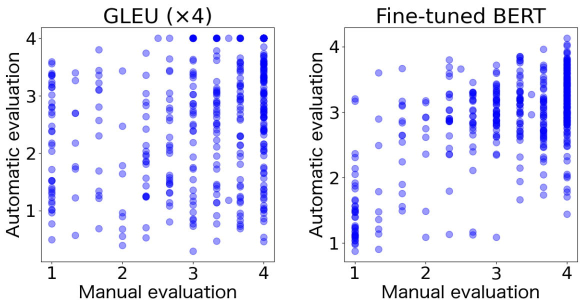

Figure 3: Sentence-level correlation between automatic and manual evaluation scores.

4.3. Analysis

To investigate the differences in the evaluations, we analyzed an example of evaluation by the two automatic evaluation methods: the quality estimation BERT model and the reference-based GLEU.

	Example 1	Example 2
S	ソンあさんが好きがっているのは推理小説です。 (Son A wants to like reading mysteries.)	堤　真一さんの演技もよかったですね。 (Shinichi Tsutsumi's acting was also good.)
C	ソンあさんが好きになっているのは推理小説です。 (Son A is becoming fond of mystery novels.)	写真一さんの演技もよかったですね。 (Photo 1's performance was also good.)
R	ソンアさんが好んでいるのは推理小説です。 (Sonna likes to read mysteries.) ソンアさんが好きなのは推理小説です。 (Sonna likes to read mysteries.)	堤　真一さんの演技もよかったですね。 (Shinichi Tsutsumi's acting was also good.) 堤　真一さんの演技もよかったですね。 (Shinichi Tsutsumi's acting was also good.)
E	GLEU, BERT, Human / 0.59, 3.30, 3.00	GLEU, BERT, Human / 3.14, 1.12, 1.00

Table 5: Successful examples of the quality estimation method. The first line shows the source text (S), the second line shows the system-corrected sentence (C), the third line shows the two reference sentences (R), and the fourth line shows the automatic and manual evaluations (E).

	Example 1	Example 2
S	社会人になりたくない原因が多いだ。 (There are many cause why I don't want to be a working adult.)	ほんとですか？ (Really?)
C	社会人になりたくない理由が多い。 (There are many reasons why I don't want to be a working adult.)	本当ですか？ (Really?)
R	社会人になりたくない原因が多い。 (There are many cause why I don't want to be a working adult.) 社会人になりたくない原因が多い。 (There are many cause why I don't want to be a working adult.)	本当ですか？ (Really?) ほんとですか？ (Really?)
E	GLEU, BERT, Human / 2.63, 1.45, 4.00	GLEU, BERT, Human / 3.41, 2.03, 4.00

Table 6: Failed examples of the quality estimation method. The first line shows the source text (S); the second line shows the system-corrected sentence (C); the third line shows the two Reference sentences (R), and the fourth line shows the automatic and manual evaluations (E).

Successful cases. Table 5 shows two examples where the quality estimation model was able to assign evaluation scores close to manual-evaluation scores.

The system-corrected sentence in Example 1 is grammatically correct, but the word "ソンあ (Sonna)", which is supposed to be a person's name, is expressed as a mixture of *katakana* and *hiragana*, resulting in a manual evaluation socre of 3.0[5]. The reference sentence offers two types of corrections for the expression "好きがっている (wants to like)"; however, because the expressions differ from the corrections made by the GEC system, the evaluation by GLEU varies considerably compared to the manual evaluation. In contrast, the quality estimation model's score was relatively close to the manual evaluation.

In Example 2, the person's name "堤　真一 (Tsutsumi Shinichi)" is corrected to "写真一 (Photo 1)" in the corrected text, and the manual evaluation score is 1 because the meaning of the original text is greatly impaired by this correction. In the reference sentence, no correction was made to the source text[6], and the output of the GEC system was superficially similar to the source text. Because GLEU calculates the score by subtracting the number of n-grams that appear in the source text but not in the reference text from the number of n-gram matches between the corrected and reference texts, it assigns a high score to the corrected text. Meanwhile, the quality estimation model captures the meaning changes between the source and reference sentences and can provide an evaluation similar to a manual evaluation.

Failed cases. We analyzed examples of evaluations in which the quality estimation model could not evaluate corrections adequately. Table 6 lists two such examples.

In Example 1, the manual evaluation score is 4.0 because the appropriate correction is made; however, the evaluation by the quality estimation model differs considerably from that performed by the manual evaluation method. Meanwhile, based on the reference sentences, GLEU can determine that the deletion of the

[5] In Japanese, katakana should be used for transliteration.

[6] In Japanese, it is incorrect to insert a space between the first and last name. TEC-JL has allowed this error by annotating it as a minor erroneous correction.

"だ (copula)" at the end of the word is an appropriate correction, and thus, GLEU is more similar to the manual evaluation than the quality estimation model.

In Example 2, the quality estimation model does not recognize the edit of "ほんと (really)" to "本当 (really)" as an appropriate correction[7]. However, GLEU can recognize the edit as an appropriate correction based on the reference sentence.

5. Conclusions

In this study, we constructed a dataset that included a manual evaluation of a holistic grammaticality-oriented approach to analyze the outputs of Japanese GEC systems. This dataset consisted of three elements: source text, corrected text, and human ratings. The source text consisted of the Lang-8 corpus, and the corrected text consisted of the outputs of the four GEC systems. The human ratings are annotated by students whose first language is Japanese.

Using the dataset, we optimized BERT directly on human ratings to create a quality estimation model. To compare the performances of the reference-less and reference-based methods, we measured the sentence-level correlation coefficients between the evaluation scores of each method and the human evaluation scores for the output of the GEC systems. The experimental results showed that the quality estimation model offered a higher correlation with manual evaluation than the reference-based method, thereby demonstrating the usefulness of a reference-less automatic evaluation method for Japanese GEC.

Future developments concerning this study include a more detailed analysis using error-type annotation and research to measure the GEC performance using reranking to select the output with the highest QE score from the outputs of multiple GEC systems.

Acknowledgments

These research results were partially obtained from the JSPS KAKENHI Grant Number 22H03651 and the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), JAPAN.

6. Bibliographical References

Asano, H., Mizumoto, T., and Inui, K. (2017). Reference-based Metrics can be Replaced with Reference-less Metrics in Evaluating Grammatical Error Correction Systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chollampatt, S. and Ng, H. T. (2018). A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5755–5762, New Orleans, Louisiana. Association for the Advancement of Artificial Intelligence.

Dale, R. and Kilgarriff, A. (2011). Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Felice, M. and Briscoe, T. (2015). Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 578–587, Denver, Colorado. Association for Computational Linguistics.

Grundkiewicz, R., Junczys-Dowmunt, M., and Gillian, E. (2015). Human Evaluation of Grammatical Error Correction Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Luong, M., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, page 1412–1421, Lisbon, Portugal. As-

[7] "本当" is pronounced as "ほんとう" and "ほんと" is a colloquial expression of "本当".

sociation for Computational Linguistics.

Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (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.

Napoles, C., Sakaguchi, K., and Tetreault, J. (2016). There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

Yoshimura, R., Kaneko, M., Kajiwara, T., and Komachi, M. (2020). SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

7. Language Resource References

Bryant, C. and Ng, H. T. (2015). How Far are We from Fully Automatic High Quality Grammatical Error Correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 697–707, Beijing, China. Association for Computational Linguistics.

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., and Tetreault, J. (2014). Predicting Grammaticality on an Ordinal Scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.

Koyama, A., Kiyuna, T., Kobayashi, K., Arai, M., and Komachi, M. (2020). Construction of an Evaluation Corpus for Grammatical Error Correction for Learners of Japanese as a Second Language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 204–211, Marseille, France. European Language Resources Association.

Mizumoto, T., Komachi, M., Nagata, M., and Matsumoto, Y. (2011). Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (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.