Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 204–211
Marseille, 11–16 May 2020

Construction of an Evaluation Corpus for Grammatical Error Correction for Learners of Japanese as a Second Language

Aomi Koyama, Tomoshige Kiyuna, Kenji Kobayashi, Mio Arai, Mamoru Komachi
Tokyo Metropolitan University
Hino City, Tokyo, Japan
{koyama-aomi@ed., kiyuna-tomoshige@ed., kobayashi-kenji1@ed., arai-mio@ed., komachi@}tmu.ac.jp

Abstract

The NAIST Lang-8 Learner Corpora (Lang-8 corpus) is one of the largest second-language learner corpora. The Lang-8 corpus is suitable as a training dataset for machine translation-based grammatical error correction systems. However, it is not suitable as an evaluation dataset because the corrected sentences sometimes include inappropriate sentences. Therefore, we created and released an evaluation corpus for correcting grammatical errors made by learners of Japanese as a Second Language (JSL). As our corpus has less noise and its annotation scheme reflects the characteristics of the dataset, it is ideal as an evaluation corpus for correcting grammatical errors in sentences written by JSL learners. In addition, we applied neural machine translation (NMT) and statistical machine translation (SMT) techniques to correct the grammar of the JSL learners' sentences and evaluated their results using our corpus. We also compared the performance of the NMT system with that of the SMT system.

Keywords: corpus construction, second-language learner corpus, Japanese grammatical error correction

1. Introduction

Grammatical error correction is the task of receiving a second-language learner sentence and outputting a sentence in which the errors were corrected. Many automatic evaluation methods in grammatical error correction systems use corrected sentences (Dahlmeier and Ng, 2012; Felice and Briscoe, 2015; Napoles et al., 2015). This requires a highly reliable evaluation corpus for an accurate evaluation (Napoles et al., 2016).

The Lang-8 corpus (Mizumoto et al., 2011) is one of the largest corpora used as a training dataset for machine translation-based grammatical error correction systems. Further, it is a corpus constructed from the revision log of Lang-8[1] and contains learner sentences and corrected sentences in nearly 80 languages. Japanese is the second largest language after English, which amounts to approximately 1.3 million sentence pairs. However, the Lang-8 corpus is not suitable as an evaluation dataset because annotators not only correct a learner's sentence, but they also sometimes write comments to the learner. Table 1 shows an example of a corrected sentence in the Lang-8 corpus. In this example, the annotator writes a comment in parentheses. These comments are useful for the learner. However, comments are noise for an evaluation dataset because automatic evaluation methods utilizing corrected sentences typically rely on the matching rate between the system output and the corrected sentences to calculate a score.

Therefore, in this study, we manually corrected the learner sentences extracted from the Lang-8 corpus using consistent rules and created a highly reliable evaluation corpus for the correction of grammatical errors in Japanese. We performed minimal edits to make the learners' sentences grammatically correct. By using our corpus as development dataset, it is possible to construct a system that performs minimal edits. In addition, in contrast to the original annotation included in the Lang-8 corpus, we created a multi-reference evaluation dataset. The Lang-8 corpus has often only one corrected sentence per learner sentence, which is not enough for evaluation. Thus, we ensured that our evaluation corpus has multiple references.

Learner's sentence	あなた**と言う**人は、天使であり、また悪魔でもあります。 (You are an angel, also a devil.)
Corrected sentence	あなた**という**人は、**私にとって**天使であり、また悪魔でもあります。（ここでの「言う」はひらがなのほうが合っています。誰もがその人を天使や悪魔だと思うわけではないので、「私にとって」を入れました） (You are an angel, also a devil for me. (“言う” (say) should be written in *hiragana* here. I added “私にとって” (for me) because not everyone thinks that person is an angel or a devil.))

Table 1: An example of a learner sentence and a corrected sentence in the Lang-8 corpus.

We also applied an NMT technique to correct the grammar in the JSL learners' sentences. Mizumoto et al. (2011) applied an SMT technique to do this; however, there are only a few studies of grammatical error correction of JSL learners' sentences using this technique. We also compared an NMT system with an SMT system.

The main contributions of this study are as follows:

- We constructed and released[2] a multi-reference corpus as the evaluation dataset for grammatical error correction of JSL learners' sentences.
- We applied an NMT technique to correct the grammar of JSL learners' sentences and evaluated the performance.

[1] Lang-8 is a social networking service where second-language learners write an article in the language they are learning and a native speaker corrects it. `http://lang-8.com`

[2] `https://forms.gle/roMnZdqd1EKWSM2D9`

Annotation rules	Learners’ sentences	Corrected sentences
L1	デザートを食べながら、**チャンペン**を飲みました。	デザートを食べながら、**シャンパン**を飲みました。 (I drank **champagne** with eating dessert.)
L2	そのあと、スポットライト**と言う**クラフトの店を見に行きました。	そのあと、スポットライト**という**クラフトの店を見に行きました。 (After that, I went to see a craft shop called Spotlight.)
G1	私のうちは四人家族です、僕と母と兄と婆ちゃんです、ちょっと狭いアパートに住んで**いる**。	私のうちは四人家族です。僕と母と兄と婆ちゃんです。ちょっと狭いアパートに住んで**います**。 (I have a family of four, me, my mother, my brother, and my grandmother. We live in a small apartment.)
G2	昨日の成績みたら、失敗しました。	昨日の成績**を**みたら、失敗していました。 (I failed because I saw yesterday’s results.)
G3	ただちょっと心配してる**ね**。。。	ただちょっと心配してる**の**。。。 (I’m just a little worried.)

Table 2: Examples of the corrected sentences based on the annotation rules in Section 3.2.

2. Related Work

2.1. Learner Corpora

“The JSL learners Parallel DataBase of Japanese writings and their translation of learners’ first language” (JPDB) (Inoue et al., 2006) is a Japanese learner corpus consisting of handwritten compositions. Each composition has the corrected sentence annotated by Japanese teachers. The NAIST Misuse Corpus (Oyama et al., 2013) is a corpus that assigns error tags to the corrected sentences in JPDB. The types of errors differ between handwritten sentences and typewritten sentences. Thus, the JPDB and NAIST Misuse Corpus are not suitable as evaluation datasets for correcting grammatical errors in typewritten sentences. Conversely, we corrected typewritten sentences to create our corpus.

Liu et al. (2018) manually performed grammatical error correction limiting error types and adding error tags in the Lang-8 corpus to study a grammatical error correction system on Japanese functional expressions. However, the authors have not released that data. We performed grammatical error correction without limiting error types and released this as a corpus.

NUCLE (Dahlmeier et al., 2013) is an annotated English learner corpus consisting of approximately 1,400 compositions written by university students in Singapore. It has sentences corrected by native English teachers corresponding to the learners’ sentences. These teachers used minimal edits to make the learners’ sentences grammatically correct. Likewise, we also used minimal edits to create the corpus for Japanese grammatical error correction.

JFLEG (Napoles et al., 2017) is an English learner corpus consisting of 747 sentences written by learners with different native languages or proficiency levels. Unlike NUCLE, the learners’ sentences are corrected by fluency edits. In addition, JFLEG is a multi-reference corpus that is corrected by four crowdsourced annotators. Similarly, we also created a multi-reference corpus of JSL learners’ sentences corrected by two or three individuals. Instead of using crowdsourcing, we created rules among the annotators to perform a consistent annotation based on minimal edits because the writings in Lang-8 are already colloquial.

2.2. Grammatical Error Correction

In previous grammatical error correction for learners of English as a second language, it is typical to limit the error types according to the parts of speech, such as articles and prepositions, and to solve the task with a classifier (De Felice and Pulman, 2008; Dahlmeier and Ng, 2011; Tajiri et al., 2012). This is because there were no publicly available large English learner corpora. Similarly, to correct grammatical errors for learners of JSL, most studies limit the target learner’s error types to mainly particles (Imaeda et al., 2003; Suzuki and Toutanova, 2006; Imamura et al., 2012).

After the emergence of a large-scale learner corpus, which is the Lang-8 corpus, it became possible to not limit the error types. Current grammatical error correction methods use SMT (Brockett et al., 2006; Junczys-Dowmunt and Grundkiewicz, 2016) and NMT (Yuan and Briscoe, 2016; Sakaguchi et al., 2017; Chollampatt and Ng, 2018; Kiyono et al., 2019) techniques extensively. The Japanese portion of the Lang-8 corpus has a wide coverage. Thus, these techniques can also be applied to JSL texts. For instance, Mizumoto et al. (2011) used the Lang-8 corpus and applied the SMT technique to correct the grammar of JSL learners’ sentences. Liu et al. (2018) applied the NMT technique to correct the grammar of Japanese sentences, but they limited the types of errors because they focused only on Japanese functional expressions. Ogawa and Yamamoto (2019) proposed a grammatical error correction system for Japanese particles based on a shallow-and-wide convolutional neural network (CNN) classification model built by training corrected sentences in the Lang-8 corpus. In Japanese grammatical error detection, Arai et al. (2019) applied the NMT technique. Moreover, they used the original annotation in the Lang-8 corpus, but they did not perform grammatical error correction because their purpose was to build an example sentence retrieval system. In this study, we apply an NMT technique to correct the grammar of JSL learners’ sentences without limiting error types and use a manually annotated corpus to evaluate the system.

	Same corrections	Different corrections
L	夜、ラベンダー**が**室内に置きます。 (Lavender will put in the room at night.)	このストーリーは、**結果がない恋**についてのちょっと悲しいストーリーです。 (This story is a little bit sad story about **love without results**.)
A	夜、ラベンダー**を**室内に置きました。 (I put lavender in the room at night.)	このストーリーは、**実らない恋**についてのちょっと悲しいストーリーです。 (This story is a slightly sad story about **unrequited love**.)
B	夜、ラベンダー**を**室内に置きます。 (I will put lavender in the room at night.)	このストーリーは、**結果のない恋**についてのちょっと悲しいストーリーです。 (This story is a slightly sad story about **love without results**.)
C	夜、ラベンダー**を**室内に置きます。 (I will put lavender in the room at night.)	このストーリーは、**成就しない恋**についてのちょっと悲しいストーリーです。 (This story is a slightly sad story about **unrequited love**.)

Table 3: Examples of a learner sentence (L) and the manual corrections of each annotator (A, B, and C). The third column shows that Annotators A and C used different surface words with the same meaning.

3. Annotation

3.1. Data

In this study, we used the Lang-8 corpus to create an evaluation corpus for grammatical error correction of JSL learners' sentences. The process for creating the evaluation corpus was as follows.

Step 1. We extracted 192,673 articles[3] of JSL learners from the Lang-8 corpus. We then randomly sampled 139 articles, which roughly amounted to 2,000 sentences, for manual annotation. These 139 articles contained 2,042 sentences.

Step 2. Three native Japanese university students made corrections until we had enough data to design the annotation rules. As a result, three annotators made corrections to 16 articles[4]. Based on these data, all the annotators discussed and decided the annotation rules.

Step 3. After the decision of the annotation rules, we assigned the remaining 123 articles[5] to two annotators, so that each sentence had two corrections. If there was any disagreement in the annotation, all the annotators discussed and made the final decision.

After creating the evaluation corpus, we compared the types of errors and their frequencies to the 16 articles annotated in step 2 to investigate trends in learner errors. Hereafter, we denote these 16 articles as the core data.

3.2. Annotation Rules

We decided to not make corrections on the sentence but on the article level. As described earlier, we performed minimum edits to make the learners' sentences grammatically correct. Moreover, when multiple interpretations were possible, the annotators corrected the sentences for each interpretation.

The annotation rules were as follows. Table 2 shows examples of the corrected sentences based on the annotation rules. Many of the articles in the Lang-8 corpus are written as if the learner writes a diary. Thus, we designed the annotation rules considering the local and global contexts dedicated to Lang-8's register (writing a blog). G2 and G3 are stylistic rules when the articles are written in such a manner.

Local rules.

L1 If non-Japanese words are written according to their original pronunciation, they are corrected to the standard notation used in Japan[6].

L2 If the subsidiary verb[7] is written in *kanji*, it is corrected to *hiragana*.

Global rules.

G1 Carry out the correction in the same article so that the formal and casual styles[8] are aligned.

G2 When there is no case particle, it is corrected if it is unnatural[9].

G3 A sentence-ending particle[10] is not corrected if the sentence does not sound unnatural[9].

3.3. Analysis

Quantitative Evaluation. To measure the inter-rater agreement rate of sentence level grammatical error detection, we calculated Fleiss' kappa for the core data. Fleiss' kappa at the sentence level was 0.72, which means the inter-rater agreement was high (Landis and Koch, 1977).

The second column in Table 3 shows an example of the same corrections in the core data. In this example, the correction of the particles is the same. The inter-rater agreement rate of correcting the particle errors was high (55.6%)[11] because the particle errors are syntactic errors

[3] Contained 1,296,114 sentences.

[4] Contained 207 sentences.

[5] Contained 1,835 sentences.

[6] We used a Web search engine to determine whether or not they were the standard notations in Japan.

[7] A verb that follows another verb that does not retain its original meaning. For example, a phrase "見ていく" (look into) consists of "見て" (look) and "いく" (go), but the core meaning of this phrase is "見て" (look) because "いく" (go) is a subsidiary verb.

[8] It is necessary to use a consistent style within the article.

[9] Each annotator judged this subjectively.

[10] A particle that appears at the end of the sentence and represents, for example, a question or an impression.

[11] The inter-rater agreement rate of detection of the particle errors was 69.4%.

Category	Subcategory	Frequency	NMT	SMT
Lexical choice	Particle	36 (10)	10	3
	Verb	13 (4)	1	0
	Noun	6 (1)	1	0
	Adjective	5 (0)	0	0
	Adverb	4 (0)	0	0
	Auxiliary verb	3 (0)	0	0
	Conjunction	3 (1)	1	0
	Adnominal	2 (0)	0	0
	Indicator	1 (1)	0	0
	Other	2 (0)	1	0
Excess or deficiency	Omitted word	22 (2)	5	3
	Redundant word	10 (0)	3	2
	Abbreviation	8 (3)	0	0
Notation	Typo	9 (0)	7	4
	Inappropriateness	8 (0)	1	0
	Transliteration	3 (0)	2	1
Verb usage	Aspect	6 (0)	1	0
	Tense	4 (4)	1	0
	Conjugation	2 (0)	1	0
	Euphony	1 (0)	1	0
Whole sentence	Formal/Casual	23 (23)	1	0
	Connection	18 (0)	1	0
	Punctuation	14 (0)	0	1
	Word order	5 (1)	1	0
	Other	2 (0)	0	0
	Total	210 (50)	39	14

Table 4: The types of errors and the frequencies of each error in the core data. The number in parentheses in the third column is the number of the errors that are not errors considering the sentence alone but are errors considering the inter-sentence context. The number in the fourth and fifth columns is the number of errors that the NMT system using the Char-Word Model and the SMT system using the Char-Char Model could correct.

captured easily using the minimal edits principle.

Table 4 shows the types of errors with their frequencies in the core data. Formal and casual style errors were not considered as errors in a single sentence because it is necessary to use a consistent style only in the entire article. Furthermore, tense errors only occur in the entire article in the core data because the tense often needs to be changed based on the inter-sentential context (Tajiri et al., 2012).

Table 5 shows the mean and variance of the edit distance (Levenshtein, 1966) between the learner's sentence and the corrected sentence for each annotator in the core data. Further, we calculated the edit distance for the corrected sentences originally given in the Lang-8 corpus[12]. It turns out that the mean and variance of the edit distance of the three annotators were almost the same. Besides this, the mean and variance of the edit distance of the original annotation were larger than that of any annotator. This is because there are no annotation rules, and comments may be inserted in the original annotation. On the other hand, in this study, the annotators made minimal edits according to consistent annotation rules; hence, the mean and variance of the edit distance were small.

	A	B	C	Original annotation
Mean	1.58	1.70	1.66	3.88
Variance	5.12	6.83	6.32	75.3

Table 5: Mean and variance of the edit distance between the learner's sentence and the corrected sentence for each annotator (A, B, and C) and the original Lang-8 annotation.

Qualitative Evaluation. The third column in Table 3 shows an example of the different corrections in the core data. The Japanese do not say “結果がない恋” (love without results). Therefore, the three annotators corrected this phrase, which has two error parts. The first part is the particle “が” (nominative case marker). The second part is the collocation between “結果” (results) and “恋” (love). The first part is clearly a grammatical error, but the second part is not necessarily an error. Therefore, Annotator B corrected only the particle, while Annotators A and C corrected both parts. As described above, the judgment on whether to correct was sometimes different for each annotator.

3.4. Comparison to NAIST Misuse Corpus

We compared the core data with the NAIST Misuse Corpus for examining the tendency of errors in typewritten and handwritten sentences of JSL learners. The particle errors accounted for a high percentage of both the core data and the NAIST Misuse Corpus. The percentage of particle errors in the core data was 17.1% and that in the NAIST Misuse Corpus was 22.1% (Oyama et al., 2013). It turns out that it is difficult for JSL learners to use particles correctly in both typewritten and handwritten sentences.

The percentages of the style and miswriting errors[13] were very different between the core data and the NAIST Misuse Corpus. The percentage of the style errors in the core data was 11.0% and that in the NAIST Misuse Corpus was 3.9%. This is because the articles recorded in the Lang-8 corpus are written as if the learner writes a diary, while the articles recorded in the NAIST Misuse Corpus are essays. The miswriting errors were idiosyncratic to the NAIST Misuse Corpus due to the handwritten compositions, which amounted to 15.7%. Meanwhile, miswriting errors did not occur in typewritten sentences, where typo errors occurred instead. The percentage of typo errors in the core data was 4.29%. This is much lower than the proportion of miswriting errors because the learner uses a conversion system when typing.

4. Experiments

4.1. Setup

We evaluated a grammatical error correction system of JSL learners' sentences using an NMT approach.

[12] When two or more corrected sentences were provided to one learner's sentence, only the first corrected sentence was used for calculating the edit distance.

[13] The miswriting error means writing nonexistent *kanji* or *hiragana* here.

Evaluation corpus	Learners' sentences	NMT system			SMT system		
		Word-Word	Char-Char	Char-Word	Word-Word	Char-Char	Char-Word
Our corpus	72.0	73.6	73.5	**73.9**	72.8	**73.9**	72.7
Lang-8 corpus	61.9	62.8	62.5	62.5	62.3	62.5	61.9

Table 6: The GLEU scores for the NMT and SMT systems using each model.

	Successful example of the Word-Word and Char-Word Models	Successful example of the Char-Char and Char-Word Models
Learner's sentence	デザートを食べながら、**チャンペン**を飲みました。	きのよるはたくやき**パーチイー**います。
Gold sentence	デザートを食べながら、**シャンパン**を飲みました。 (I drank **champagne** with eating dessert.)	きのうのよるはたこやき**パーティー**にいました。 (I was at a *takoyaki* **party** last night.)
Word-Word Model	デザートを食べながら、**シャンパン**を飲みました。	きのうはたくやき**パーチイー**います。
Char-Char Model	デザートを食べながら、**チャンペン**を飲みました。	きのうはたくやき**パーティー**があります。
Char-Word Model	デザートを食べながら、**シャンパン**を飲みました。	きのうは、たくやき**パーティー**をします。

Table 7: Examples of outputs from the NMT system using each model.

Corpus. The training dataset was composed of JSL learners' sentences in the Lang-8 corpus, excluding sentences used to create our corpus. Pre-processing was performed on the training dataset. Following Mizumoto et al. (2011), we removed sentence pairs whose edit distance between the learner's sentence and the corrected sentence was 7 or more. We also removed the sentence pairs whose length of the learner's sentence was more than 100 characters, and those which had no edit. As a result of this pre-processing, 200,439 sentence pairs were removed from the training dataset. The remaining 1,093,633 sentence pairs were used as the training dataset.

The corpus created in Section 3.1 was used as the development and evaluation dataset. We used 806 sentences for development and 663 sentences for evaluation. In addition, the Lang-8's original annotation[14] corresponding to our corpus was also used for the evaluation dataset for comparison with our corpus.

Tokenization. We tested the following three tokenization models. Japanese is a language where there are no spaces between words. Therefore, we usually perform word segmentation using a morphological analyzer as a pre-processing step. However, if a sentence contains grammatical errors, it is difficult to tokenize correctly. To alleviate this problem, we used the Char-Char Model and Char-Word Model proposed in Mizumoto et al. (2011).

1. Word-Word Model

   In this model, we tokenized both the learner sentence and the corrected sentence at the word level. This was the baseline model.

2. Char-Char Model

   In this model, we tokenized both the learner sentence and the correction sentence at the character level. By tokenizing at character level, this model was not affected by word segmentation errors.

3. Char-Word Model

   In this model, we tokenized the learner sentence at the character level. On the other hand, we tokenized a corrected sentence at the word level. There is a high possibility that word segmentation can be performed correctly in the corrected sentence because the corrected sentence does not contain grammatical errors. It is expected to be able to use word level information, while reducing the effects of word segmentation errors.

NMT system. We used a CNN based method (Chollampatt and Ng, 2018) for the grammatical error correction system using the NMT technique. We used the implementation[15] published by Chollampatt and Ng (2018). Both the source and target embeddings were of 500 dimensions. Each encoder and decoder was made up of seven convolutional layers, with a convolution window width of three. The output of each encoder and decoder layer was of 1,024 dimensions. We used MeCab[16] (ver.0.996) using UniDic[17] (ver.2.2.0) as a dictionary for word segmentation. Furthermore, we used Byte Pair Encoding (BPE) (Sennrich et al., 2016) for subword processing of rare words, and the vocabulary size was 30,000 words. In addition, we converted full-width to half-width characters using mojimoji[18] (ver.0.0.9).

SMT system. As a comparative experiment, we evaluated a grammatical error correction system of JSL learners' sentences using the SMT technique. We used Moses[19] (Koehn et al., 2007) as a method for the SMT toolkit and set distortion-limit to the value -1. We also used GIZA++[20] (Och and Ney, 2003) as the word alignment tool. Following Mizumoto et al. (2011), we created a word 3-gram language model and character 5-gram language model from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) using KenLM

[14] Parentheses and characters in the parentheses were removed using a regular expression.

[15] `https://github.com/nusnlp/mlconvgec2018`

[16] `https://taku910.github.io/mecab`

[17] `https://unidic.ninjal.ac.jp`

[18] `https://pypi.org/project/mojimoji`

[19] `http://statmt.org/moses`

[20] `https://github.com/moses-smt/giza-pp`

	TP	FP	FN	Precision	Recall	$F_{0.5}$	Insertion	Deletion	Substitution
NMT system	39	187	171	17.3	18.6	17.5	135	15	76
SMT system	14	24	196	36.8	6.67	**19.3**	12	7	19

Table 8: Analysis of the NMT system using the Char-Word Model and the SMT system using the Char-Char Model in the core data.

(Heafield, 2011). We used the word 3-gram language model in the Word-Word Model and Char-Word Model. We also used the character 5-gram language model in the Char-Char Model. In addition, in the SMT system, we added sentence pairs that copied the corrected sentences in the training dataset to the learner sentence as new training data to reduce unknown words. We performed word segmentation and converted full-width to half-width characters in the same way as for the NMT system. However, BPE was not used in the SMT system.

Metric. We used the generalized language evaluation understanding metric (GLEU) (Napoles et al., 2015) to evaluate the performance of each grammatical error correction system. We used 4-grams when calculating the GLEU score. In the NMT system, training was terminated at the epoch in which the best GLEU score was achieved in the development dataset. The maximum number of epochs was 100. The parameters for the SMT system were adjusted to maximize BLEU (Papineni et al., 2002) using MERT (Och, 2003) for the development dataset.

4.2. Results

Table 6 shows the GLEU scores of the NMT and SMT systems using each model. We also calculated the GLEU score of the learners' sentences. In the NMT system, the GLEU score of the NMT system using the Char-Word Model was the highest. On the other hand, in the SMT system, the GLEU score of the SMT system using the Char-Char Model was the highest. For any output, changing the evaluation corpus from our corpus to Lang-8 corpus reduced the GLEU score by 10 or more points because of remaining comments that could not be removed using a regular expression.

The GLEU score of the NMT system using the Word-Word Model was higher than that of the NMT system using the Char-Char Model. This is because rare words were split into characters by BPE, and the NMT system using the Word-Word Model could use word-level information while reducing the effects of word segmentation errors. Table 7 shows outputs from the NMT system using each model. In the second column, the Word-Word Model and Char-Word Model correct "チャンペン" (champagne) correctly. In the third column, the Char-Char Model and Char-Word Model correct "パーチイー" (party) correctly. In many cases, the Char-Word Model could correct a part that either or both the Word-Word Model and Char-Char Model could correct.

We compared the NMT system using the Char-Word Model with the SMT system using the Char-Char Model because they had the highest GLEU scores for each system. The fourth and fifth columns in Table 4 show the number of errors that the NMT system using the Char-Word Model and the SMT system using the Char-Char Model could correct in the core data. The NMT system had a high accuracy of correcting particles and typo errors. Therefore, it turns out that a CNN-based method is effective for errors that can be corrected with only the local context (Chollampatt and Ng, 2018). In contrast, both the NMT and SMT systems could hardly correct errors that needed to be considered in context, for example, abbreviation or formal and casual style errors. Table 8 shows a analysis of the NMT system using the Char-Word Model and the SMT system using the Char-Char Model in the core data. The number of true positives (TP) in the NMT system was larger than that in the SMT system. On the other hand, the number of false positives (FP) in the NMT system was considerably larger than that in the SMT system. In other words, the NMT system changed many points that did not need to be changed. In addition, it turns out that the number of corrections in the SMT system was smaller than that in the NMT system. This is because we tuned the BLEU score using MERT, and the SMT system learned parameter weights that disabled nearly all correction attempts (Junczys-Dowmunt and Grundkiewicz, 2014). As a result, the precision and $F_{0.5}$ of the NMT system are lower than thoes of the SMT system.

5. Conclusions

We created and released a highly reliable evaluation corpus for a grammatical error correction system of JSL learners' sentences. Unlike the Lang-8 corpus, our corpus is suitable as an evaluation dataset for grammatical error correction of JSL learners' sentences. Lang-8's original annotation contains annotator's comments that are noise for evaluation. In contrast, our evaluation corpus does not contain such comments. In addition, in many cases, only one corrected sentence is provided per learner sentence in the Lang-8 corpus. However, we ensured that our evaluation corpus has multiple references.

We applied an NMT technique to correct the grammar of JSL learners' sentences and compared the NMT system with an SMT system. In addition, we tested different granularities of tokenization proposed in Mizumoto et al. (2011) in the NMT system. As a result, we confirmed that the Char-Word Model in the NMT system and the Char-Char Model in the SMT system recorded the highest GLEU.

We compared the types of errors and their frequencies in the core data. As a future work, we will examine the types of errors and their frequencies in full sentences of our evaluation corpus and add error tags (Oyama et al., 2013).

6. Acknowledgements

We would like to thank the organizers of Lang-8 web for providing the text data. We would also like to thank the anonymous reviewers for their valuable feedback.

7. Bibliographical References

Arai, M., Kaneko, M., and Komachi, M. (2019). Grammatical-Error-Aware Incorrect Example Retrieval System for Learners of Japanese as a Second Language. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 296–305, Florence, Italy, August. Association for Computational Linguistics.

Brockett, C., Dolan, W. B., and Gamon, M. (2006). Correcting ESL Errors Using Phrasal SMT Techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia, July. Association for Computational Linguistics.

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

Dahlmeier, D. and Ng, H. T. (2011). Grammatical Error Correction with Alternating Structure Optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 915–923, Portland, Oregon, USA, June. Association for Computational Linguistics.

Dahlmeier, D. and Ng, H. T. (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, June. Association for Computational Linguistics.

De Felice, R. and Pulman, S. G. (2008). A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 169–176, Manchester, UK, August. COLING 2008 Organizing Committee.

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, May–June. 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, July. Association for Computational Linguistics.

Imaeda, K., Kawai, A., Ishikawa, Y., Nagata, R., and Masui, F. (2003). Error detection and correction of case particles in Japanese Learner's composition. *The Special Interest Group Technical Reports of IPSJ*, 2003(13):39–46.

Imamura, K., Saito, K., Sadamitsu, K., and Nishikawa, H. (2012). Grammar Error Correction Using Pseudo-Error Sentences and Domain Adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 388–392, Jeju Island, Korea, July. Association for Computational Linguistics.

Junczys-Dowmunt, M. and Grundkiewicz, R. (2014). The AMU System in the CoNLL-2014 Shared Task: Grammatical Error Correction by Data-Intensive and Feature-Rich Statistical Machine Translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 25–33, Baltimore, Maryland, June. Association for Computational Linguistics.

Junczys-Dowmunt, M. and Grundkiewicz, R. (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, November. Association for Computational Linguistics.

Kiyono, S., Suzuki, J., Mita, M., Mizumoto, T., and Inui, K. (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*, pages 1236–1242, Hong Kong, China, November. 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 *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. 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.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Liu, J., Cheng, F., Wang, Y., Shindo, H., and Matsumoto, Y. (2018). Automatic Error Correction on Japanese Functional Expressions Using Character-based Neural Machine Translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, pages 394–403, Hong Kong, December. Association for Computational Linguistics.

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, November. Asian Federation of Natural Language Processing.

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, pages 588–593, Beijing, China, July. 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, November. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Ogawa, Y. and Yamamoto, K. (2019). Japanese Particle Error Correction employing Classification Model. In *Proceedings of the 2019 International Conference on Asian Language Processing*, pages 23–28, Shanghai, China, November. Institute of Electrical and Electronics Engineers.

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, July. Association for Computational Linguistics.

Sakaguchi, K., Post, M., and Van Durme, B. (2017). Grammatical Error Correction with Neural Reinforcement Learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 366–372, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Suzuki, H. and Toutanova, K. (2006). Learning to Predict Case Markers in Japanese. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1056, Sydney, Australia, July. Association for Computational Linguistics.

Tajiri, T., Komachi, M., and Matsumoto, Y. (2012). Tense and Aspect Error Correction for ESL Learners Using Global Context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 198–202, Jeju Island, Korea, July. Association for Computational Linguistics.

Yuan, Z. and Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California, June. Association for Computational Linguistics.

8. Language Resource References

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

Inoue, M., Usami, Y., Narita, T., and Yarimizu, K. (2006). *For Various Uses of the Composition Parallel Database*. National Institute for Japanese Language and Linguistics, 1st edition.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Napoles, C., Sakaguchi, K., and Tetreault, J. (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*, pages 229–234, Valencia, Spain, April. Association for Computational Linguistics.

Oyama, H., Komachi, M., and Matsumoto, Y. (2013). Towards Automatic Error Type Classification of Japanese Language Learners' Writings. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*, pages 163–172, Taipei, Taiwan, November. Department of English, National Chengchi University.