Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 340–345
Marseille, 11–16 May 2020

Creating Corpora for Research in Feedback Comment Generation

Ryo Nagata, Kentaro Inui, Shin'ichiro Ishikawa
Konan University/JST, Presto, Tohoku University/RIKEN AIP, Kobe University
8-9-1 Okamoto, Higashinada-ku, Kobe, Hyogo 658-8501, Japan
6-6-05 Aoba, Aramaki, Aoba-ku, Sendai, 980-8579, Japan
1-2-1 Tsurukabuto, Nada-ku, Kobe, Hyogo 657-8501, Japan
nagata-lrec2019 @ ml.hyogo-u.ac.jp., inui@ecei.tohoku.ac.jp, aiskwshin@gmail.com

Abstract

In this paper, we report on datasets that we created for research in feedback comment generation — a task of automatically generating feedback comments such as a hint or an explanatory note for writing learning. There has been almost no such corpus open to the public and accordingly there has been a very limited amount of work on this task. In this paper, we first discuss the principle and guidelines for feedback comment annotation. Then, we describe two corpora that we have manually annotated with feedback comments (approximately 50,000 general comments and 6,700 on preposition use). A part of the annotation results is now available on the web, which will facilitate research in feedback comment generation.

Keywords: Feedback comment generation, Writing learning, Grammatical error correction

1. Introduction

This paper reports on datasets that we have created for research in feedback comment generation. Feedback comment generation is a task of generating feedback comments for a given text (hereafter, *essay*) automatically. A feedback comment is a hint or an explanatory note for the writer (typically, a language learner) that helps them improve their writing skill. As shown in Fig. 1 and Fig. 2, it is typically a comment on erroneous, unnatural, or problematic words in a given text so that the writer can understand why the present form does not match with the underlying rule.

The purpose of this corpus creation is to facilitate research in feedback comment generation. It will augment grammatical error correction as language learning assistance if one can automatically generate natural and effective feedback comments with grammatical error correction results.

Unfortunately, however, there has been a very limited amount of work on feedback comment generation as Sect. 2. will describe. This is largely due to the fact that there had been no publicly available corpus annotated with feedback comments. Corpora annotated with feedback comments will likely facilitate the research in this domain just as the common datasets in grammatical error correction such as the CoNLL shared-task datasets (Ng et al., 2013; Ng et al., 2014) did. Recently, Nagata (2019) has released a dataset consisting of learner corpora manually annotated with feedback comments on preposition use; their target corpora are the two existing corpora — the written essays in ICNALE (Ishikawa, 2013) and KJ (Nagata et al., 2011).

In this work, we extended the dataset by manually annotating the two learner corpora with feedback comments in general and those on preposition use (hereafter, general and preposition feedback comments, respectively). We also extended guidelines for this annotation. We annotated approximately 3,000 and 2,400 essays with 50,000 general and 6,700 preposition feedback comments, respectively. We released a part of these annotated corpora to the public on the web[1]. We are now planning to organize a feedback comment generation shared task in the near future[2] (this is why only a part of the data is available at present).

2. Related Work

More and more learner corpora have become available to the public. Without doubt, they have greatly contributed to the recent improvements in Natural Language Processing (NLP) techniques related to learner language including grammatical error correction and automated essay scoring;

In the beginning, raw learner corpora made their appearance to the public. Examples are: ICLE (Granger, 1993), NICT JLE (Izumi et al., 2004), and ICNALE, to name a few. They are crucial sources of information about learner language.

Since around 2000, annotated learner corpora have become available to the public. Now, a wide variety of them are available. Examples are corpora annotated with grammatical errors and/or spelling errors as in a part of NICT JLE, CLC FCE (Yannakoudakis et al., 2011), KJ (Nagata et al., 2011), and the CoNLL shared-task datasets (Ng et al., 2013; Ng et al., 2014). Related to these are parallel corpora where the original and its corrected sentences are paired as in the Lang-8 learner corpus (Lang-8) (Mizumoto et al., 2012) and JFLEG (Napoles et al., 2017). Another direction is syntactic annotation as in the work by Nagata and Sakaguchi (2016) and Berzak et al. (2016).

In contrast, there have been almost no publicly available corpora annotated with feedback comments despite the fact that they are essential for research in feedback comment generation. An exception is the dataset reported in the work (Nagata, 2019). The reason why datasets for feedback comment generation are rare is that it is highly costly

[1]The corpus data are available together with the guidelines at ICNALE Learner Essays with Feedback Comments (`https://www.gsk.or.jp/en/catalog/gsk2019-b`) and Konan-JIEM Learner Corpus Sixth Edition (`https://www.gsk.or.jp/en/catalog/gsk2019-a`)

[2]The details of the shared task are available at `https://fcg.sharedtask.org/`

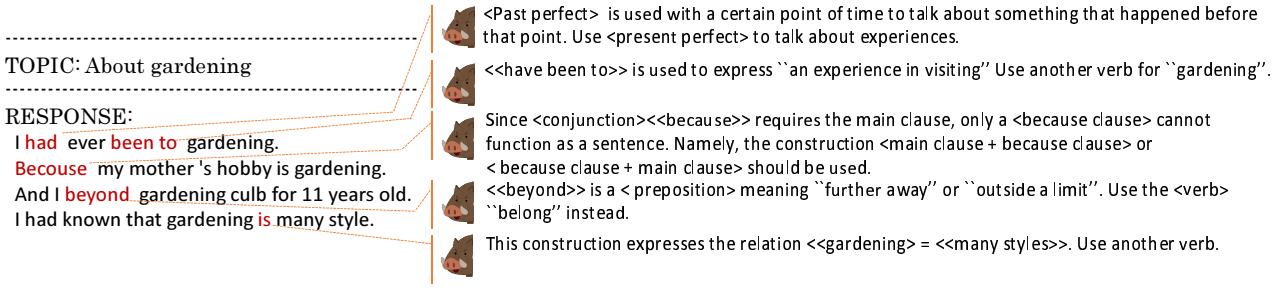

Figure 1: Example of General Feedback Comments.

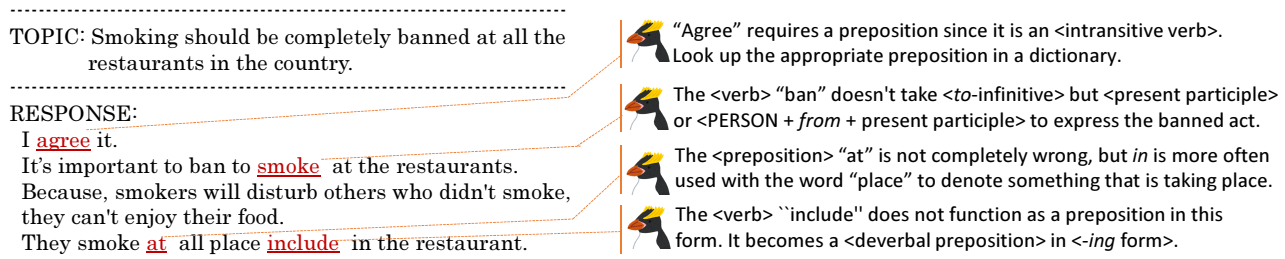

Figure 2: Example of Feedback Comments on Preposition Use.

and time-consuming to create such a corpus. In addition, it is not trivial at all to decide what sort of feedback comment to give. In English language teaching research, there is a great amount of work on how to correct grammatical errors as in the work by (Robb et al., 1986; Ferris and Roberts, 2001) but little work on what information we should provide as feedback comments and how; as far as we know, Bitchener et al. (2005) show that error correction can significantly improve learners' writing accuracy levels when combined with feedback messages[3].

Because of this data limitation, there has been only a small amount of work on feedback comment generation. Some researchers as Kakegawa et al. (2000) and McCoy et al. (1996) made an attempt to develop handcrafted rules to diagnose errors. Nagata et al. (2014) proposed to use automatically extracted case frames to correct preposition errors with explanations. These approaches encounter the difficulty of covering a wide variety of errors. More recently, Lai and Chang (2019) proposed a method that uses grammatical error correction and templates to generate detailed comments. Nagata (2019) reports on performance of a neural retrieval-based method on their dataset.

3. Corpus Design and Guidelines

3.1. Target Corpus

As already mentioned in Sect. 1., ICNALE and KJ are our base corpora. ICNALE has several suitable properties for this task. In particular, its essay topics are well-controlled, which enables us to simulate a situation common to language teaching and learning where all learners write on the same topic as in writing exercises in class and writing tasks in proficiency tests. It would be interesting to see how well we can generate feedback comments under this condition. In ICNALE, all essays are written on two common topics: (a) *It is important for college students to have a part-time job*; (b) *Smoking should be completely banned at all the restaurants in the country* (both are argumentative essays). In addition to this, the size of ICNALE is relatively large, amounting to 5,600 essays (approximately 1,300,000 tokens; 200 to 300 tokens per essay on average). The writers are college students (including graduate students) from 10 countries and regions in Asia. Their proficiency levels range from A2 to B2+ in the CERF level.

In contrast, the other corpus, KJ, is much smaller (only 233 essays). However, it has nice properties that other corpora do not; it is fully annotated with spelling and grammatical errors, Part-Of-Speech (POS), and phrase structures[4], which might be useful for feedback comment generation.

3.2. Annotation Principle

It is not straightforward at all to decide what sort of feedback comment to give. The first choice would be comments on grammatical errors, but they can be on other things such as organization and mechanics to improve one's writing skill. They can even be praise to motivate the writer.

To get an idea of what sort of feedback comments we should annotate, we performed two trial sessions where two annotators freely added feedback comments to 20 essays (10 per session) sampled out of KJ and ICNALE. One was a professional annotator who had a good command of English; she had experience in English writing teaching for two years and also in English syntactic annotation for more than ten years. The other was the first author. They used

[3]Note that in their work, human teachers did error correction in written form and provided the learners with feedback messages orally.

[4]However, we did not use the information for this annotation to examine what we should annotate from scratch.

the commenting function in the MS-Word software for annotation. At the end of each of the sessions, they had a discussion session to make a draft annotation guideline.
As a result, they agreed on the following annotation principle:

Principle: Annotate the given essay with feedback comments most relevant to the proficiency level of the writer.

It would be useless to give the writer feedback comments that are too difficult to understand or that have not yet been introduced to the writer. Besides, too many feedback comments would be unrealistic from learners' point of view (also from generation point of view). Ideally, feedback comments will facilitate learning writing if they fit the writer's proficiency level and if they are adequate in amount. The principle states that the annotator should focus on the most relevant feedback comments. Of course, what is relevant would vary depending on the writer's proficiency and the annotator. Considering this, we let the annotator estimate the writer's proficiency and then decide what is relevant to the writer. Because of this nature of the annotation principle, annotation results can involve feedback comments on a wide variety of writing techniques including grammatical errors, lexical choices, organization, and mechanics as Sect. 4. will show.
In addition, we include feedback comments on preposition use in the annotation. It is much clearer to determine where and what to annotate in this case. Accordingly, we target all preposition errors in the corpora (plus, annotators may add positive feedback comments where they want to). We annotate preposition feedback comments separately from general ones. Thus, the annotated feedback comments on preposition use partly overlap with the general ones (but not completely).
Before annotation, we have to choose in which language we will create feedback comments. The choice should be between English and the writer's native language. We choose Japanese, one of the learners' native languages, for the following reasons: (1) Beginner to intermediate learners may have difficulty in understanding feedback comments in English when working on writing exercises; (2) It will likely be more technically challenging and interesting to generate feedback comments on English in a different language; (3) it would be too costly to create feedback comments in all the native languages, and accordingly, we have to choose one. To augment accessibility, a part of annotated feedback comments are translated into English in the dataset. Also, feedback comments are directly written in English in some essays.

3.3. Annotation Procedure and Guidelines

Our basic annotation procedure for general feedback comments is as follows:

1. Read the entire essay first before annotation
2. Determine what sort of feedback comments are most relevant to the writer
3. Annotate the given essay with about five to ten feedback comments[5] based on 2
4. After annotation, double-check the results
5. Revise the results (if necessary)

For preposition feedback comments, the second step is replaced by *Determine where preposition errors exist and other places that require feedback comments.* because all preposition errors are the target of the annotation.
During the trial annotations, we made two special annotation symbols for grammatical terms ($<$, $>$) and citations ($<<$, $>>$). Grammatical terms are tagged inside $<$ and $>$ as in $<$*intransitive verb*$>$. With this, one can make links to corresponding grammatical items in a grammar book, for example, as an additional source of information for the user[6]. Citations are used to denote that the word(s) inside the symbols is cited from the commented sentence as in $<<$*because*$>>$, which might be useful for feedback comment generation.
This is the big picture of our corpus design and annotation guidelines. The complete details are in the guidelines accompanying the dataset.

4. Annotation and the Data

For general feedback comments, we hired twelve annotators. They were either item writers/editors/ex-editors of English learning materials, or raters for English proficiency tests. For preposition feedback comments, we hired two professional annotators who had a good command of English. Both of them had experience in English syntactic annotation for more than ten years. One of them had also two years of English writing teaching experience.
After hiring them, we performed another two trial sessions for both general and preposition feedback comment annotations to fine-tune the corresponding annotation guidelines. We sampled 200 essays out of ICNALE and KJ. We assigned them to two annotators both for general and preposition feedback comments in the same manner as in the first two trials. When the annotators finished the trials, they once again checked the whole results. In the meantime, they revised the annotation guidelines for general and preposition feedback comments.
Finally, they started to annotate the entire corpora. For general feedback comments, two of the twelve annotators were assigned to each essay, resulting in two versions of annotation for each essay. In contrast, for preposition feedback comment annotation, only one annotator was assigned to each essay. As preprocessing, the essays were split into sentences, and in turn tokenized using the Stanford Statistical Natural Language Parser (ver.2.0.3) (de Marneffe et al., 2006).
Table 1 and Table 2 show the statistics on the essays that have been annotated so far. To obtain the statistics, we developed a tool to transform the MS-Word format into a TSV

[5]Note that the target essay is assumed to consist of 200 to 300 tokens just as in ICNALE.

[6]We have also created a grammar database with a list of grammar items and their explanations. However, it is not included in the dataset due to copyright issues.

Corpus	ICNALE		KJ	
Annotation	# 1	# 2	# 1	# 2
Number of essays	2,541	2,300	233	233
Number of sentences	38,214	34,667	3,236	3,236
Number of tokens	644,625	581,533	30,802	30,802
Number of comments	22,899	19,405	2,005	2,037
Comment rate (per sentence)	0.60	0.56	0.62	0.63

Table 1: Statistics on General Comment Annotation Results.

Corpus	ICNALE	KJ
Number of essays	2,077	233
Number of sentences	31,292	3,236
Number of tokens	524,973	30,802
Number of comments	6,148	538
Comment rate (per sentence)	0.20	0.17

Table 2: Statistics on Preposition Comment Annotation Results.

format (learner sentence, feedback comment, offset indicating to which word(s) the feedback comment applies). We also released the tool to the public on the web[7].

5. Looking into the Data

We looked into the results to develop an understanding of what had been annotated. We sampled 20 pairs of essays from the two versions of general feedback comment annotation. One set of the 20 contained 175 feedback comments, and the other contained 182. We then manually compared them and found that 78 feedback comments from each set were the same or almost the same comments in terms of their content. This corresponds to a Szymkiewicz-Simpson coefficient of 44.6%, which shows a mild agreement between the two versions of feedback comments despite the great degree of freedom in theory. In other words, feedback comment annotation converges to some extent, which is a nice property for training methods for generating feedback comments and evaluating them.

We further looked into the types of the general feedback comments. We sampled 169 feedback comments from the two versions and then manually classified them according to their types. Table 3 shows the results. It reveals that while most general feedback comments are about grammatical errors, they can also be about things such as lexical choice, organization, and, mechanics.

Feedback comment about grammatical errors range over a wide variety of error types. Major error types include preposition errors concerning transitive/intransitive verbs:

> <<Agree>> is an <intransitive verb> when used to express "to be in favor of something" which requires a <preposition>.

and

> <<Control>> is a <transitive verb> and does not need a <preposition>.,

article errors:

> <<Book>> is a countable <noun> and should not be used in the ¡singular form without an article¿. Use the <bare plural form> of the noun in the expression "to like reading books" to refer to multiple and unspecified books.,

and tense errors:

> Use the same <tense> to describe the event as it took place at the same time as the subject "thought".

The former two are relatively easy to deal with because they involve relatively narrow contexts. In contrast, the latter is much more difficult because it requires understanding of a wider context and/or the intention of the writer.

Approximately 16% of general feedback comments are about lexical choice as in:

> Non-smoking <<seats>> are used for trains and airplanes. Use "tables" for restaurants.

This type of feedback comments include knowledge about word meaning and/or collocation. This implies that it would require a certain kind of grounding to dictionaries to generate feedback comments of this type.

Organization and mechanics are much less frequent than grammatical errors. The former is about text structures and relations between sentences as in:

> <<On the other hand>> should be used to express comparison. Use other expressions for consequence.

The latter is about writing rules such as spelling, capitalization, and punctuation as in:

> Keep in mind that <<However>> is normally followed by a comma when used to argue against what is discussed in the preceding part.

[7] RIKEN Wex (https://www.gsk.or.jp/en/catalog/gsk2019-c)

Type	Ratio (%)
Grammatical error	61.5
Lexical choice	16.0
Organization	10.7
Mechanics	10.1
Other	0.02

Table 3: Types of General Feedback Comments.

and

> Pay attention to the spelling when you write a loanword in English.

It will be interesting to see how we can automatically generate feedback comments about lexical choice, organization, and mechanics on top of grammatical errors.

6. Conclusions

In this paper, we reported on datasets that we created for research in feedback comment generation. First, we discussed the principle and guidelines for annotation. Based on them, we manually annotated ICNALE and KJ with general feedback comments and those on preposition use. We looked into the created data, showing their types and tendencies. To be precise, general feedback comments range over grammatical errors, lexical choice, organization, and mechanics. We further discussed which types of feedback comments were expected to be easy or difficult to generate. A part of the created dataset is publicly available on the web.

In future work, we will develop methods for automatically generating feedback comments using the dataset. We are also planning to organize a shared task on feedback comment generation.

Acknowledgments

This work was partly supported by JST, PRESTO Grant Number JPMJPR1758, Japan.

7. Bibliographical References

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). Universal dependencies for learner English. In *Proc. of 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746.

Bitchener, J., Young, S., and Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14(3):191–205.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proc. of 5th International Conference on Language Resources and Evaluation*, pages 449–445.

Ferris, D. and Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, 10(3):161–184.

Granger, S. (1993). The international corpus of learner English. In *English language corpora: Design, analysis and exploitation*, pages 57–69. Rodopi.

Ishikawa, S., (2013). *The ICNALE and Sophisticated Contrastive Interlanguage Analysis of Asian learners of English*, pages 91–118. Kobe University, Kobe.

Izumi, E., Saiga, T., Supnithi, T., Uchimoto, K., and Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of The Computer, the Internet and Management*, 12(2):119–125.

Kakegawa, J., Kanda, H., Fujioka, E., Itami, M., and Itoh, K. (2000). Diagnostic processing of Japanese for computer-assisted second language learning. In *Proc. of 38th Annual Meeting of the Association for Computational Linguistics*, pages 537–546.

Lai, Y.-H. and Chang, J. (2019). TellMeWhy: Learning to explain corrective feedback for second language learners. In *Proc. of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 235–240.

McCoy, K. F., Pennington, C. A., and Suri, L. Z. (1996). English error correction: A syntactic user model based on principled "mal-rule" scoring. In *Proc. of 5th International Conference on User Modeling*, pages 69–66.

Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M., and Matsumoto, Y. (2012). The effect of learner corpus size in grammatical error correction of ESL writings. In *Proc. of 24th International Conference on Computational Linguistics*, pages 863–872.

Nagata, R. and Sakaguchi, K. (2016). Phrase structure annotation and parsing for learner English. In *Proc. of 54th Annual Meeting of the Association for Computational Linguistics*, pages 1837–1847.

Nagata, R., Whittaker, E., and Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proc. of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1219.

Nagata, R., Vilenius, M., and Whittaker, E. (2014). Correcting preposition errors in learner English using error case frames and feedback messages. In *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–764.

Nagata, R. (2019). Toward a task of feedback comment generation for writing learning. In *Proc. of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3197–3206.

Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proc. of 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–234.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The CoNLL-2013 shared task on

grammatical error correction. In *Proc. 17th Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.

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 *Proc. 18th Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Robb, T., Ross, S., and Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL QUARTERY*, 20(1):83–93.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proc. of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.