A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English

Ryo Nagata
Hyogo University of Teacher Education
6731494, Japan
rnagata@hyogo-u.ac.jp

Atsuo Kawai
Mie University
5148507, Japan
kawai@ai.info.mie-u.ac.jp

Koichiro Morihiro
Hyogo University of Teacher Education
6731494, Japan
mori@hyogo-u.ac.jp

Naoki Isu
Mie University
5148507, Japan
isu@ai.info.mie-u.ac.jp

Abstract

This paper proposes a method for detecting errors in article usage and singular plural usage based on the mass count distinction. First, it learns decision lists from training data generated automatically to distinguish mass and count nouns. Then, in order to improve its performance, it is augmented by feedback that is obtained from the writing of learners. Finally, it detects errors by applying rules to the mass count distinction. Experiments show that it achieves a recall of 0.71 and a precision of 0.72 and outperforms other methods used for comparison when augmented by feedback.

1 Introduction

Although several researchers (Kawai et al., 1984; McCoy et al., 1996; Schneider and McCoy, 1998; Tschichold et al., 1997) have shown that rule-based methods are effective to detecting grammatical errors in the writing of learners of English, it has been pointed out that it is hard to write rules for detecting errors concerning the articles and singular plural usage. To be precise, it is hard to write rules for distinguishing mass and count nouns which are particularly important in detecting these errors (Kawai et al., 1984). The major reason for this is that whether a noun is a mass noun or a count noun greatly depends on its meaning or its surrounding context (refer to Allan (1980) and Bond (2005) for details of the mass count distinction).

The above errors are very common among Japanese learners of English (Kawai et al., 1984; Izumi et al., 2003). This is perhaps because the Japanese language does not have a mass count distinction system similar to that of English. Thus, it is favorable for error detection systems aiming at Japanese learners to be capable of detecting these errors. In other words, such systems need to somehow distinguish mass and count nouns.

This paper proposes a method for distinguishing mass and count nouns in context to complement the conventional rules for detecting grammatical errors. In this method, first, training data, which consist of instances of mass and count nouns, are automatically generated from a corpus. Then, decision lists for distinguishing mass and count nouns are learned from the training data. Finally, the decision lists are used with the conventional rules to detect the target errors.

The proposed method requires a corpus to learn decision lists for distinguishing mass and count nouns. General corpora such as newspaper articles can be used for the purpose. However, a drawback to it is that there are differences in character between general corpora and the writing of non-native learners of English (Granger, 1998; Chodorow and Leacock, 2000). For instance, Chodorow and Leacock (2000) point out that the word *concentrate* is usually used as a noun in a general corpus whereas it is a verb 91% of the time in essays written by non-native learners of English. Consequently, the differences affect the performance of the proposed method.

In order to reduce the drawback, the proposed method is augmented by feedback; it takes as feedback learners' essays whose errors are corrected by a teacher of English (hereafter, referred to as the feedback corpus). In essence, the feedback corpus could be added to a general corpus to generate training data. Or, ideally training data could be generated only from the feedback corpus just as

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 241–248,
Sydney, July 2006.

from a general corpus. However, this causes a serious problem in practice since the size of the feedback corpus is normally far smaller than that of a general corpus. To make it practical, this paper discusses the problem and explores its solution.

The rest of this paper is structured as follows. Section 2 describes the method for detecting the target errors based on the mass count distinction. Section 3 explains how the method is augmented by feedback. Section 4 discusses experiments conducted to evaluate the proposed method.

2 Method for detecting the target errors

2.1 Generating training data

First, instances of the target noun that head their noun phrase (NP) are collected from a corpus with their surrounding words. This can be simply done by an existing chunker or parser.

Then, the collected instances are tagged with mass or count by the following tagging rules. For example, the underlined *chicken*:

... are a lot of *chicken*s in the roost ...

is tagged as

... are a lot of *chickens*/count in the roost ...

because it is in plural form.

We have made tagging rules based on linguistic knowledge (Huddleston and Pullum, 2002). Figure 1 and Table 1 represent the tagging rules. Figure 1 shows the framework of the tagging rules. Each node in Figure 1 represents a question applied to the instance in question. For example, the root node reads "Is the instance in question plural?". Each leaf represents a result of the classification. For example, if the answer is *yes* at the root node, the instance in question is tagged with count. Otherwise, the question at the lower node is applied and so on. The tagging rules do not classify instances as mass or count in some cases. These unclassified instances are tagged with the symbol "?". Unfortunately, they cannot readily be included in training data. For simplicity of implementation, they are excluded from training data[1].

Note that the tagging rules can be used only for generating training data. They cannot be used to distinguish mass and count nouns in the writing of learners of English for the purpose of detecting the target errors since they are based on the articles and the distinction between singular and plural.

Finally, the tagged instances are stored in a file with their surrounding words. Each line of it consists of one of the tagged instances and its surrounding words as in the above *chicken* example.

2.2 Learning Decision Lists

In the proposed method, decision lists are used for distinguishing mass and count nouns. One of the reasons for the use of decision lists is that they have been shown to be effective to the word sense disambiguation task and the mass count distinction is highly related to word sense as we will see in this section. Another reason is that rules for distinguishing mass and count nouns are observable in decision lists, which helps understand and improve the proposed method.

A decision list consists of a set of rules. Each rule matches the template as follows:

If *a condition is true*, then *a decision*. (1)

To define the template in the proposed method, let us have a look at the following two examples:

1. I read the *paper*.
2. The *paper* is made of hemp pulp.

The underlined *paper*s in both sentences cannot simply be classified as mass or count by the tagging rules presented in Section 2.1 because both are singular and modified by the definite article. Nevertheless, we can tell that the former is a count noun and the latter is a mass noun from the contexts. This suggests that the mass count distinction is often determined by words surrounding the target noun. In example 1, we can tell that the *paper* refers to something that can be read such as a newspaper or a scientific paper from *read*, and therefore it is a count noun. Likewise, in example 2, we can tell that the *paper* refers to a certain substance from *made* and *pulp*, and therefore it is a mass noun.

Taking this observation into account, we define the template based on words surrounding the target noun. To formalize the template, we will use a random variable MC that takes either $mass$ or $count$ to denote that the target noun is a mass noun or a count noun, respectively. We will also use w and C to denote a word and a certain context around the target noun, respectively. We define

[1] According to experiments we have conducted, approximately 30% of instances are tagged with "?" on average. It is highly possible that performance of the proposed method will improve if these instances are included in the training data.

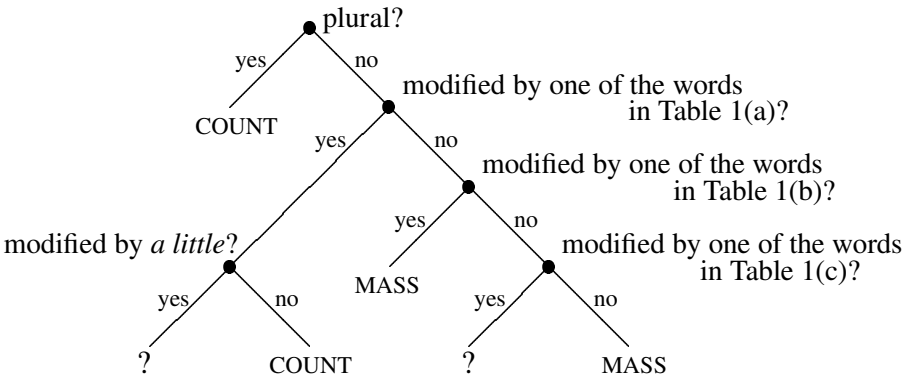

Figure 1: Framework of the tagging rules

Table 1: Words used in the tagging rules

(a)	(b)	(c)
the indefinite article	much	*the definite article*
another	less	*demonstrative adjectives*
one	enough	*possessive adjectives*
each	sufficient	*interrogative adjectives*
–	–	*quantifiers*
–	–	*'s genitives*

three types of C: np, $-k$, and $+k$ that denote the contexts consisting of the noun phrase that the target noun heads, k words to the left of the noun phrase, and k words to its right, respectively. Then the template is formalized by:

If word w appears in context C of the target noun, then it is distinguished as MC.

Hereafter, to keep the notation simple, it will be abbreviated to

$$w_C \rightarrow MC. \tag{2}$$

Now rules that match the template can be obtained from the training data. All we need to do is to collect words in C from the training data. Here, the words in Table 1 are excluded. Also, function words (except prepositions), cardinal and quasi-cardinal numerals, and the target noun are excluded. All words are reduced to their morphological stem and converted entirely to lower case when collected. For example, the following tagged instance:

She ate fried *chicken*/mass for dinner.

would give a set of rules that match the template:

$$eat_{-3} \rightarrow mass$$
$$fry_{np} \rightarrow mass$$
$$for_{+3} \rightarrow mass$$
$$dinner_{+3} \rightarrow mass$$

for the target noun *chicken* when $k = 3$.

In addition, a default rule is defined. It is based on the target noun itself and used when no other applicable rules are found in the decision list for the target noun. It is defined by

$$t \rightarrow MC_{\text{major}} \tag{3}$$

where t and MC_{major} denote the target noun and the majority of MC in the training data, respectively. Equation (3) reads "If the target noun appears, then it is distinguished by the majority".

The log-likelihood ratio (Yarowsky, 1995) decides in which order rules are applied to the target noun in novel context. It is defined by[2]

$$log\frac{p(MC|w_C)}{p(\overline{MC}|w_C)} \tag{4}$$

where $\overline{MC}$ is the exclusive event of MC and $p(MC|w_C)$ is the probability that the target noun is used as MC when w appears in the context C.

It is important to exercise some care in estimating $p(MC|w_C)$. In principle, we could simply

[2]For the default rule, the log-likelihood ratio is defined by replacing w_C and MC with t and MC_{major}, respectively.

count the number of times that w appears in the context C of the target noun used as MC in the training data. However, this estimate can be unreliable, when w does not appear often in the context. To solve this problem, using a smoothing parameter α (Yarowsky, 1996), $p(MC|w_C)$ is estimated by[3]

$$p(MC|w_C) = \frac{f(w_C, MC) + \alpha}{f(w_C) + m\alpha} \quad (5)$$

where $f(w_C)$ and $f(w_C, MC)$ are occurrences of w appearing in C and those in C of the target noun used as MC, respectively. The constant m is the number of possible classes, that is, $m = 2$ ($mass$ or $count$) in our case, and introduced to satisfy $p(MC|w_C) + p(\overline{MC}|w_C) = 1$. In this paper, α is set to 1.

Rules in a decision list are sorted in descending order by the log-likelihood ratio. They are tested on the target noun in novel context in this order. Rules sorted below the default rule are discarded[4] because they are never used as we will see in Section 2.3.

Table 2 shows part of a decision list for the target noun *chicken* that was learned from a subset of the BNC (British National Corpus) (Burnard, 1995). Note that the rules are divided into two columns for the purpose of illustration in Table 2; in practice, they are merged into one.

Table 2: Rules in a decision list

Mass		Count	
w_C	LLR	w_C	LLR
$piece_{-3}$	1.49	$count_{-3}$	1.49
$fish_{-3}$	1.28	$peck_{+3}$	1.32
$dish_{-3}$	1.23	pig_{np}	1.23
$skin_{+3}$	1.23	run_{-3}	1.23
$serve_{+3}$	1.18	egg_{np}	1.18

target noun: *chicken*, $k = 3$
LLR (Log-Likelihood Ratio)

On one hand, we associate the words in the left half with food or cooking. On the other hand, we associate those in the right half with animals or birds. From this observation, we can say that *chicken* in the sense of an animal or a bird is a count noun but a mass noun when referring to food or cooking, which agrees with the knowledge presented in previous work (Ostler and Atkins, 1991).

2.3 Distinguishing mass and count nouns

To distinguish the target noun in novel context, each rule in the decision list is tested on it in the sorted order until the first applicable one is found. It is distinguished according to the first applicable one. Ties are broken by the rules below.

It should be noted that rules sorted below the default rule are never used because the default rule is always applicable to the target noun. This is the reason why rules sorted below the default rule are discarded as mentioned in Section 2.2.

2.4 Detecting the target errors

The target errors are detected by the following three steps. Rules in each step are examined on each target noun in the target text.

In the first step, any mass noun in plural form is detected as an error[5]. If an error is detected in this step, the rest of the steps are not applied.

In the second step, errors are detected by the rules described in Table 3. The symbol "×" in Table 3 denotes that the combination of the corresponding row and column is erroneous. For example, the fifth row denotes that singular and plural count nouns modified by *much* are erroneous. The symbol "–" denotes that no error can be detected by the table. If one of the rules in Table 3 is applied to the target noun, the third step is not applied.

In the third step, errors are detected by the rules described in Table 4. The symbols "×" and "–" are the same as in Table 3.

In addition, the indefinite article that modifies other than the head noun is judged to be erroneous

Table 3: Detection rules (i)

	Count		Mass
Pattern	Sing.	Pl.	Sing.
{another, each, one}	–	×	×
{all, enough, sufficient}	×	–	–
{much}	×	×	–
{that, this}	–	×	–
{few, many, several}	×	–	×
{these, those}	×	–	×
{various, numerous}	×	–	×
cardinal numbers exc. one	×	–	×

[3]The probability for the default rule is estimated just as the log-likelihood ratio for the default rule above.

[4]It depends on the target noun how many rules are discarded.

[5]Mass nouns can be used in plural in some cases. However, they are rare especially in the writing of learners of English.

Table 4: Detection rules (ii)

	Singular			Plural		
	a/an	the	ϕ	a/an	the	ϕ
Mass	×	–	–	×	×	×
Count	–	–	×	×	–	–

(e.g., *an expensive). Likewise, the definite article that modifies other than the head noun or adjective is judged to be erroneous (e.g., *the them). Also, we have made exceptions to the rules. The following combinations are excluded from the detection in the second and third steps: head nouns modified by interrogative adjectives (e.g., what), possessive adjectives (e.g., my), 's genitives, "some", "any", or "no".

3 Feedback-augmented method

As mentioned in Section 1, the proposed method takes the feedback corpus[6] as feedback to improve its performance. In essence, decision lists could be learned from a corpus consisting of a general corpus and the feedback corpus. However, since the size of the feedback corpus is normally far smaller than that of general corpora, so is the effect of the feedback corpus on $p(MC|w_C)$. This means that the feedback corpus hardly has effect on the performance.

Instead, $p(MC|w_C)$ can be estimated by interpolating the probabilities estimated from the feedback corpus and the general corpus according to confidences of their estimates. It is favorable that the interpolated probability approaches to the probability estimated from the feedback corpus as its confidence increases; the more confident its estimate is, the more effect it has on the interpolated probability. Here, confidence c of ratio p is measured by the reciprocal of variance of the ratio (Tanaka, 1977). Variance is calculated by

$$\frac{p(1-p)}{n} \tag{6}$$

where n denotes the number of samples used for calculating the ratio. Therefore, confidence of the estimate of the conditional probability used in the proposed method is measured by

$$c = \frac{f(w_C)}{p(MC|w_C)(1-p(MC|w_C))}. \tag{7}$$

[6]The feedback corpus refers to learners' essays whose errors are corrected as mentioned in Section 1.

To formalize the interpolated probability, we will use the symbols p_{fb}, p_g, c_{fb}, and c_g to denote the conditional probabilities estimated from the feedback corpus and the general corpus, and their confidences, respectively. Then, the interpolated probability p_i is estimated by[7]

$$p_i = \begin{cases} p_g + \frac{c_{fb}}{c_g}(p_{fb} - p_g), & c_{fb} < c_g \\ p_{fb}, & c_{fb} \geq c_g \end{cases}. \tag{8}$$

In Equation (8), the effect of p_{fb} on p_i becomes large as its confidence increases. It should also be noted that when its confidence exceeds that of p_g, the general corpus is no longer used in the interpolated probability.

A problem that arises in Equation (8) is that p_{fb} hardly has effect on p_i when a much larger general corpus is used than the feedback corpus even if p_{fb} is estimated with a sufficient confidence. For example, p_{fb} estimated from 100 samples, which are a relatively large number for estimating a probability, hardly has effect on p_i when p_g is estimated from 10000 samples; roughly, p_{fb} has a $1/100$ effect of p_g on p_i.

One way to prevent this is to limit the effect of c_g to some extent. It can be realized by taking the log of c_g in Equation (8). That is, the interpolated probability is estimated by

$$p_i = \begin{cases} p_g + \frac{c_{fb}}{\log c_g}(p_{fb} - p_g), & c_{fb} < \log c_g \\ p_{fb}, & c_{fb} \geq \log c_g \end{cases}. \tag{9}$$

It is arguable what base of the log should be used. In this paper, it is set to 2 so that the effect of p_g on the interpolated probability becomes large when the confidence of the estimate of the conditional probability estimated from the feedback corpus is small (that is, when there is little data in the feedback corpus for the estimate)[8].

In summary, Equation (9) interpolates between the conditional probabilities estimated from the feedback corpus and the general corpus in the feedback-augmented method. The interpolated probability is then used to calculate the log-likelihood ratio. Doing so, the proposed method takes the feedback corpus as feedback to improve its performance.

[7]In general, the interpolated probability needs to be normalized to satisfy $\sum p_i = 1$. In our case, however, it is always satisfied without normalization since $p_{fb}(MC|w_C) + p_{fb}(\overline{MC}|w_C) = 1$ and $p_g(MC|w_C) + p_g(\overline{MC}|w_C) = 1$ are satisfied.

[8]We tested several bases in the experiments and found there were little difference in performance between them.

4 Experiments

4.1 Experimental Conditions

A set of essays[9] written by Japanese learners of English was used as the target essays in the experiments. It consisted of 47 essays (3180 words) on the topic *traveling*. A native speaker of English who was a professional rewriter of English recognized 105 target errors in it.

The written part of the British National Corpus (BNC) (Burnard, 1995) was used to learn decision lists. Sentences the OAK system[10], which was used to extract NPs from the corpus, failed to analyze were excluded. After these operations, the size of the corpus approximately amounted to 80 million words. Hereafter, the corpus will be referred to as the BNC.

As another corpus, the English concept explication in the EDR English-Japanese Bilingual dictionary and the EDR corpus (1993) were used; it will be referred to as the EDR corpus, hereafter. Its size amounted to about 3 million words.

Performance of the proposed method was evaluated by recall and precision. Recall is defined by

$$\frac{\text{No. of target errors detected correctly}}{\text{No. of target errors in the target essays}}. \quad (10)$$

Precision is defined by

$$\frac{\text{No. of target errors detected correctly}}{\text{No. of detected errors}}. \quad (11)$$

4.2 Experimental Procedures

First, decision lists for each target noun in the target essays were learned from the BNC[11]. To extract noun phrases and their head nouns, the OAK system was used. An optimal value for k (window size of context) was estimated as follows. For 25 nouns shown in (Huddleston and Pullum, 2002) as examples of nouns used as both mass and count nouns, accuracy on the BNC was calculated using ten-fold cross validation. As a result of setting small ($k = 3$), medium ($k = 10$), and large ($k = 50$) window sizes, it turned out that $k = 3$ maximized the average accuracy. Following this result, $k = 3$ was selected in the experiments.

Second, the target nouns were distinguished whether they were mass or count by the learned decision lists, and then the target errors were detected by applying the detection rules to the mass count distinction. As a preprocessing, spelling errors were corrected using a spell checker. The results of the detection were compared to those done by the native-speaker of English. From the comparison, recall and precision were calculated.

Then, the feedback-augmented method was evaluated on the same target essays. Each target essay in turn was left out, and all the remaining target essays were used as a feedback corpus. The target errors in the left-out essay were detected using the feedback-augmented method. The results of all 47 detections were integrated into one to calculate overall performance. This way of feedback can be regarded as that one uses revised essays previously written in a class to detect errors in essays on the same topic written in other classes.

Finally, the above two methods were compared with their seven variants shown in Table 5. "DL" in Table 5 refers to the nine decision list based methods (the above two methods and their seven variants). The words in brackets denote the corpora used to learn decision lists; the symbol "+FB" means that the feedback corpus was simply added to the general corpus. The subscripts fb_1 and fb_2 indicate that the feedback was done by using Equation (8) and Equation (9), respectively.

In addition to the seven variants, two kinds of earlier method were used for comparison. One was one (Kawai et al., 1984) of the rule-based methods. It judges singular head nouns with no determiner to be erroneous since missing articles are most common in the writing of Japanese learners of English. In the experiments, this was implemented by treating all nouns as count nouns and applying the same detection rules as in the proposed method to the countability.

The other was a web-based method (Lapata and Keller, 2005)[12] for generating articles. It retrieves web counts for queries consisting of two words preceding the NP that the target noun head, one of the articles ($\{$a/an, the, $\phi\}$), and the core NP to generate articles. All queries are performed as exact matches using quotation marks and submitted to the Google search engine in lower case. For example, in the case of "*She is good student.", it retrieves web counts for "she is a good student",

[9] http://www.eng.ritsumei.ac.jp/lcorpus/.

[10] OAK System Homepage: http://nlp.cs.nyu.edu/oak/.

[11] If no instance of the target noun is found in the general corpora (and also in the feedback corpus in case of the feedback-augmented method), the target noun is ignored in the error detection procedure.

[12] There are other statistical methods that can be used for comparison including Lee (2004) and Minnen (2000). Lapata and Keller (2005) report that the web-based method is the best performing article generation method.

"she is the good student", and "she is good student". Then, it generates the article that maximizes the web counts. We extended it to make it capable of detecting our target errors. First, the singular/plural distinction was taken into account in the queries (e.g., "she is a good students", "she is the good students", and "she is good students" in addition to the above three queries). The one(s) that maximized the web counts was judged to be correct; the rest were judged to be erroneous. Second, if determiners other than the articles modify head nouns, only the distinction between singular and plural was taken into account (e.g., "he has some book" vs "he has some books"). In the case of "much/many", the target noun in singular form modified by "much" and that in plural form modified by "many" were compared (e.g., "he has much furniture" vs "he has many furnitures). Finally, some rules were used to detect literal errors. For example, plural head nouns modified by "this" were judged to be erroneous.

4.3 Experimental Results and Discussion

Table 5 shows the experimental results. "Rule-based" and "Web-based" in Table 5 refer to the rule-based method and the web-based method, respectively. The other symbols are as already explained in Section 4.2.

As we can see from Table 5, all the decision list based methods outperform the earlier methods. The rule-based method treated all nouns as count nouns, and thus it did not work well at all on mass nouns. This caused a lot of false-positives and false-negatives. The web-based method suffered a lot from other errors than the target errors since it implicitly assumed that there were no errors except the target errors. Contrary to this assumption, not only did the target essays contain the target errors but also other errors since they were written by Japanese learners of English. This indicate that the queries often contained the other errors when web counts were retrieved. These errors made the web counts useless, and thus it did not perform well. By contrast, the decision list based methods did because they distinguished mass and count nouns by one of the words around the target noun that was most likely to be effective according to the log-likelihood ratio[13]; the best performing decision list based method (DL_{fb2} (EDR)) is significantly superior to the best performing[14] non-decision list based method (Web-based) in both recall and precision at the 99% confidence level.

Table 5: Experimental results

Method	Recall	Precision
DL (BNC)	0.66	0.65
DL (BNC+FB)	0.66	0.65
DL_{fb1} (BNC)	0.66	0.65
DL_{fb2} (BNC)	0.69	0.70
DL (EDR)	0.70	0.68
DL (EDR+FB)	0.71	0.69
DL_{fb1} (EDR)	0.71	0.70
DL_{fb2} (EDR)	0.71	0.72
DL (FB)	0.43	0.76
Rule-based	0.59	0.39
Web-based	0.49	0.53

Table 5 also shows that the feedback-augmented methods benefit from feedback. Only an exception is "DL_{fb1} (BNC)". The reason is that the size of BNC is far larger than that of the feedback corpus and thus it did not affect the performance. This also explains that simply adding the feedback corpus to the general corpus achieved little or no improvement as "DL (EDR+FB)" and "DL (BNC+FB)" show. Unlike these, both "DL_{fb2} (BNC)" and "DL_{fb2} (EDR)" benefit from feedback since the effect of the general corpus is limited to some extent by the log function in Equation (9). Because of this, both benefit from feedback despite the differences in size between the feedback corpus and the general corpus.

Although the experimental results have shown that the feedback-augmented method is effective to detecting the target errors in the writing of Japanese learners of English, even the best performing method (DL_{fb2} (EDR)) made 30 false-negatives and 29 false-positives. About 70% of the false-negatives were errors that required other sources of information than the mass count distinction to be detected. For example, extra definite articles (e.g., *the traveling) cannot be detected even if the correct mass count distinction is given. Thus, only a little improvement is expected in recall however much feedback corpus data become available. On the other hand, most of the

[13] Indeed, words around the target noun were effective. The default rules were used about 60% and 30% of the time in "DL (EDR)" and "DL (BNC)", respectively; when only the default rules were used, "DL (EDR)" ("DL (BNC)") achieved 0.66 (0.56) in recall and 0.58 (0.53) in precision.

[14] "Best performing" here means best performing in terms of F-measure.

false-positives were due to the decision lists themselves. Considering this, it is highly possible that precision will improve as the size of the feedback corpus increases.

5 Conclusions

This paper has proposed a feedback-augmented method for distinguishing mass and count nouns to complement the conventional rules for detecting grammatical errors. The experiments have shown that the proposed method detected 71% of the target errors in the writing of Japanese learners of English with a precision of 72% when it was augmented by feedback. From the results, we conclude that the feedback-augmented method is effective to detecting errors concerning the articles and singular plural usage in the writing of Japanese learners of English.

Although it is not taken into account in this paper, the feedback corpus contains further useful information. For example, we can obtain training data consisting of instances of errors by comparing the feedback corpus with its original corpus. Also, comparing it with the results of detection, we can know performance of each rule used in the detection, which make it possible to increase or decrease their log-likelihood ratios according to their performance. We will investigate how to exploit these sources of information in future work.

Acknowledgments

The authors would like to thank Sekine Satoshi who has developed the OAK System. The authors also would like to thank three anonymous reviewers for their useful comments on this paper.

References

K. Allan. 1980. Nouns and countability. *J. Linguistic Society of America*, 56(3):541–567.

F. Bond. 2005. *Translating the Untranslatable*. CSLI publications, Stanford.

L. Burnard. 1995. *Users Reference Guide for the British National Corpus. version 1.0*. Oxford University Computing Services, Oxford.

M. Chodorow and C. Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proc. of 1st Meeting of the North America Chapter of ACL*, pages 140–147.

Japan electronic dictionary research institute ltd. 1993. *EDR electronic dictionary specifications guide*. Japan electronic dictionary research institute ltd, Tokyo.

S. Granger. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. P. Cowie, editor, *Phraseology: theory, analysis, and applications*, pages 145–160. Clarendon Press.

R. Huddleston and G.K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge.

E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *Proc. of 41st Annual Meeting of ACL*, pages 145–148.

A. Kawai, K. Sugihara, and N. Sugie. 1984. ASPEC-I: An error detection system for English composition. *IPSJ Journal (in Japanese)*, 25(6):1072–1079.

M. Lapata and F. Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–31.

J. Lee. 2004. Automatic article restoration. In *Proc. of the Human Language Technology Conference of the North American Chapter of ACL*, pages 31–36.

K.F. McCoy, C.A. Pennington, and L.Z. Suri. 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.

G. Minnen, F. Bond, and A. Copestake. 2000. Memory-based learning for article generation. In *Proc. of CoNLL-2000 and LLL-2000 workshop*, pages 43–48.

N. Ostler and B.T.S Atkins. 1991. Predictable meaning shift: Some linguistic properties of lexical implication rules. In *Proc. of 1st SIGLEX Workshop on Lexical Semantics and Knowledge Representation*, pages 87–100.

D. Schneider and K.F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *Proc. of 17th International Conference on Computational Linguistics*, pages 1198–1205.

Y. Tanaka. 1977. *Psychological methods (in Japanese)*. University of Tokyo Press.

C. Tschichold, F. Bodmer, E. Cornu, F. Grosjean, L. Grosjean, N. Kübler, N. Lévy, and C. Tschumi. 1997. Developing a new grammar checker for English as a second language. In *Proc. of the From Research to Commercial Applications Workshop*, pages 7–12.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of 33rd Annual Meeting of ACL*, pages 189–196.

D. Yarowsky. 1996. *Homograph Disambiguation in Speech Synthesis*. Springer-Verlag.