RECOGNIZING ARTICLE ERRORS BASED ON THE
THREE HEAD WORDS

Ryo Nagata, Fumito Masui, Atsuo Kawai, and Naoki Isu
Department of Information Engineering, Mie University
1515, Kamihama, Tsu, 514-8507, Japan

ABSTRACT
In this paper, we present a statistical model for detecting article errors, which Japanese learners of English often make in English writing. The model detects article errors based on probabilities estimated looking at the three head words - the head verb, the preposition, and the head noun in a corpus. Unfortunately, simply looking at the three head words causes the data sparseness problem. To abate the problem, we apply the backed-off estimate to estimate the probabilities. Experiments show that the performance ($F$-measure=0.70) of the model is better than that of other methods. Apart from the performance, the model has two advantages: (i) Rules for detecting article errors are automatically generated as conditional probabilities once a corpus is given. (ii) Recall and precision rates of the model are adjustable.

KEYWORDS
Article errors, Corpus, Statistics, English writing, Japanese learners of English

1. INTRODUCTION
One of the big problems Japanese learners of English have in English writing is the use of articles. It is difficult for Japanese learners of English to use articles properly, perhaps because the Japanese language does not have an article system similar to that of English. Consequently, Japanese learners of English often make article errors in English writing (e.g., "*I am student.").

Correlated with this is the use of mass and count nouns. The Japanese language also has no or little conception of mass and count nouns. Because of that reason, mass nouns are often used improperly as count nouns in the writing of Japanese learners of English and vice versa (e.g., "*informations"). Hereafter, to keep the notation simple, errors concerning the use of mass and count nouns will be referred to as article error as well as article error itself.

Article errors can also be a big problem for Japanese teachers of English. Japanese teachers of English have to correct a lot of article errors in writing such as essays because Japanese learners of English often make article errors as we have just mentioned. In some cases, even Japanese teachers of English need to look up in a dictionary or check with examples when correcting article errors because of the variety of article usage. It is laborious to correct article errors looking up in a dictionary or checking with examples.

In view of this fact, several methods for detecting article errors have been proposed in the past. One way of detecting article errors is to use rules (Kawai et al., 1984; Schneider and McCoy, 19989). For example, a rule for detecting the dropping of articles would be as follows: If a count noun is used as singular without articles, then detect it as an article error. Another way is to use statistics. A statistical model for detecting article errors has been proposed in previous work (Nagata et al., 2004). The model is based on probabilities estimated from a corpus that a noun occurs with each article. Unlike the rule-based methods, rules for detecting article errors are automatically generated as probabilities once a corpus is given.

A common problem to both the rule-based methods and the statistical model is that the performances are rather low. One of the major reasons for that is due to the variety of article usage. The variety of article usage makes it difficult to construct rules for detecting article errors. For example, since some nouns are used as both mass and count nouns, it is difficult to construct rules for distinguishing mass and count nouns precisely (Kawai et al., 1984). Besides, simply adding rules causes significant performance problems (kawai et al.,
1985; Schneider and McCoy, 1998). In the statistical model, the situation is similar. The statistical model cannot cover all the variety.

In this paper, we propose a method for improving the statistical model. In this method, probabilities are estimated looking at the three head words — the VP (Verb Phrase) head, the preposition and the NP (Noun Phrase) head — instead of looking at just the NP head (that is adopted in the previous statistical model (Nagata et al., 2004)). Unfortunately, simply looking at the three head words causes the data sparseness problem. To abate the data sparseness problem, we apply the backed-off estimate (Collins and Brooks, 1995) to the new statistical model.

In the next section, we present the new statistical model based on the three head words. In Section 3, we then describe the approach taken to reducing the data sparseness problem. In Section 4, we describe the results of the experiments conducted to evaluate the model.

2. MODEL BASED ON THE THREE HEAD WORDS

2.1 Article class

In English, there are two kinds of article: the indefinite article and the definite article. In addition to these articles, we admit that no article is one of the articles and will denote it by \( \phi \).

In combination with the articles and singular/plural, we have six classes: three (the articles) multiplied by two (the contrast between singular and plural). Of the six classes, the combination of the indefinite article and plural is ungrammatical in English. Nevertheless, we define it, because it appears in the writing of Japanese learners of English.

Hereafter, to keep the notation simple, we will denote the six classes by article class or the symbol \( I_i \) where \( i \) ranges from 1 to 6 and corresponds to the six classes (Table 1). Table 1 shows the correspondence between \( I_i \) and the six classes.

<table>
<thead>
<tr>
<th></th>
<th>Indefinite article</th>
<th>Definite article</th>
<th>( \phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td>( I_1 )</td>
<td>( I_2 )</td>
<td>( I_3 )</td>
</tr>
<tr>
<td>Plural</td>
<td>( I_4 )</td>
<td>( I_5 )</td>
<td>( I_6 )</td>
</tr>
</tbody>
</table>

2.2 Model based on the three head word

As we have already mentioned, probabilities in the new statistical model are based on the three head words. Before estimating such probabilities, for each NP in the corpus, the head verb and the preposition dominating the NP are extracted along with the head noun itself and its article class \( I_i \). All head words are reduced to their morphological stem and converted entirely to lower case, when extracted. For example, the phrase:

\( (VP \text{ laughed (PP at (NP a funny joke))}) \)

would give the quadruple:

\( I_1 \\text{ laugh at joke} \)

The elements of this quadruple will be referred to random variables \( L, V, PREP \) and \( N \), hereafter. In the above example, \( L = I_1 \), \( V = \text{ laugh} \), \( PREP = \text{ at} \), and \( N = \text{ joke} \).

As for phrases that have no preposition or head verb, the symbol \( \varepsilon \) is used. For example, the phrase:

\( (VP \text{ have (NP a book)}) \)
would give the quadruple:

\[ l_1 \text{ have } \epsilon \text{ book} \]

Likewise, the underlined phrase:

(NP (NP a book) (PP on (NP science)))

would give the quadruple:

\[ l_3 \text{ on } \text{ science} \]

We will also use the symbol \( f \) to denote the number of occurrences of a particular tuple in the corpus. For example, \( f(l_1, \text{laught}, \text{at}, \text{joke}) \) is the number of occurrences of the triple \((\text{laugh}, \text{at}, \text{joke})\) where the noun \text{joke} falls into the article class \( l_1 \). Occurrences of lower order tuples can be defined in the same manner.

For example, \( f(l_1, \text{PREP = at}, N = \text{joke}) \) is the number of occurrences of the pair \((\text{at}, \text{joke})\) and any value of \( V \) where the noun \text{joke} falls into the article class \( l_1 \).

Then, the probability of the article class variable \( L \) being \( l_i \) is estimated as follows:

\[
p(l_i \mid v, prep, n) = \frac{f(l_i, v, prep, n)}{\sum_{m=1}^{6} f(l_m, v, prep, n)}.
\]  

Eq. (1) could serve as a criterion of detecting article errors. For instance, if the head noun of a particular triple rarely falls into the article class \( l_i \), that is, \( p(l_i \mid v, prep, n) \) is relatively low, we can say that it might be an article error in the writing of the Japanese learners of English on the assumption that Japanese learners of English do not know the rare use of articles and therefore mistakenly use it.

To measure how relatively low \( p(l_i \mid v, prep, n) \) is, we formalize

\[
s(l_i, v, prep, n) = \frac{p(l_i \mid v, prep, n)}{\max_{1 \leq m \leq 6} p(l_m \mid v, prep, n)}.
\]  

Eq. (2) expresses how relatively low \( p(l_i \mid v, prep, n) \) is as the ratio of \( p(l_i \mid v, prep, n) \) to \( p(l_m \mid v, prep, n) \) that gives the maximum value of the conditional probability.

A particular triple that satisfies

\[
s(l_i, v, prep, n) < \theta
\]

is detected as an article error. Here, \( \theta \) is a threshold determining whether the triple in question is correct or not. (How to set \( \theta \) will soon be described in the next subsection). When \( L=L_5 \), false-positives often occur based on just \( s(l_5, v, prep, n) < \theta \) because the article class \( l_5 \) appears relatively infrequently compared to the other classes as shown in Table 2. "Freq." in Table 2 corresponds to the relative frequencies of \( L \) estimated from a corpus (We will mention the detail of the corpus in Subsection 4.1). To reduce the false-positives, when \( L=L_5 \), triples are detected as article errors only if \( s(l_2, v, prep, n) < \theta \) and \( s(l_6, v, prep, n) < \theta \) are satisfied along with \( s(l_5, v, prep, n) < \theta \). The former is used for determining whether or not \( n \) with the definite article is correct. Likewise, the latter is used for determining whether or not \( n \) being plural is correct.

In the case of NPs that have neither head verb nor preposition, the conditional probability is estimated by

\[
p(l_i \mid n) = \frac{f(l_i, n)}{\sum_{m=1}^{6} f(l_m, n)}
\]

instead of eq. (1). Likewise, eq. (2) and eq. (3) are defined based on \((l_i, n)\) (e.g., \( s(l_i, n) \)).


Table 2. Relative frequencies of $L$

<table>
<thead>
<tr>
<th>$L$</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
<th>$l_4$</th>
<th>$l_5$</th>
<th>$l_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>0.25</td>
<td>0.20</td>
<td>0.32</td>
<td>0</td>
<td>0.06</td>
<td>0.16</td>
</tr>
</tbody>
</table>

2.3 Determining the threshold

The threshold $\theta$ in eq. (3) significantly affects the performance of the model. Applying eq. (3) with $\theta$ ($0 \leq \theta \leq 1$) to a large set of revised essays written by Japanese learners of English would give the optimal threshold. In general, however, it is impractical because there is no such set of essays opened to the public. Because of that reason, we use a corpus to determine $\theta$. We can reasonably assume that there are only a few article errors in a corpus such as English-language newspapers because they are usually written by native speakers of English and proofread sufficiently. Therefore it is highly possible that triples detected as article errors by applying eq. (3) to them are not actually article errors. Based on this observation, we can calculate error rates by

$$e(\theta) = \frac{\text{No. of triples detected as errors}}{\text{No. of triples examined}}$$

In Determining $\theta$, we put emphasis on precision rather than recall because false-positives confuse learners; this conception is applied in previous work (Chodorow and Leacock, 2002). To put emphasis on precision, we determine $\theta$ that gives a small $e(\theta)$ (e.g., $e(\theta) = 0.1$).

3. REDUCING THE DATA SPARSENESS PROBLEM

Unfortunately, the data sparseness problem disables eq. (3) from detecting article errors. A triple may appear in the writing of the Japanese learners of English which have never been seen in the corpus used for estimating the conditional probabilities. In that case, the conditional probability is undefined because the denominator of eq. (1) ($\sum_{m=1}^{6} f(l_m, v, prep, n)$) is 0. It would be the case even if very large corpora were used for the estimate. Even if $\sum_{m=1}^{6} f(l_m, v, prep, n) > 0$, it may still be too low to make an accurate estimate. It takes at least five occurrences to see a triple $(v, prep, n)$ with each $l_i$ except for $l_4$.

To abate the data sparseness problem, we apply the backed-off estimate described in (Collins and Brooks, 1995) to the model. In the backed-off estimate, triples whose occurrences are insufficient are estimated based on lower order tuples. Here, the problem is that we have three possible pairs $(v, prep)$, $(v, n)$, $(prep, n)$ when backing off from triples $(v, prep, n)$. Intuitively, the head noun is most important taking into account the fact that articles modify the head noun. After the head noun, we need to choose either the VP head or the preposition. In choosing between the VP head and the preposition, we make an assumption that the preposition has a greater affect on the use of articles because there are a lot of examples of prepositions that affect article class (e.g., ``It was an accident," and ``It was by accident"). Thus, we choose the pair $(prep, n)$ to back-off from the triple $(v, prep, n)$.

In summary, the backed-off estimate we adopt in this paper is as follows:

Stage 1. If $\sum_{m=1}^{6} f(l_m, v, prep, n) > c_1$,

$$p(l_i \mid v, prep, n) = \frac{f(l_i, v, prep, n)}{\sum_{m=1}^{6} f(l_m, v, prep, n)},$$

Stage 2. If $\sum_{m=1}^{6} f(l_m, prep, n) > c_2$
\[
p(l_i | v, prep, n) = \frac{f(l_i, prep, n)}{\sum_{m=1}^{6} f(l_m, prep, n)},
\]

Stage 3. If \(\sum_{m=1}^{6} f(l_m, n) > c_3\)

\[
p(l_i | v, prep, n) = \frac{f(l_i, n)}{\sum_{m=1}^{6} f(l_m, n)},
\]

else abort the estimation and detection.

Here, \(c_1, c_2,\) and \(c_3\) are thresholds called cut off frequency that determine whether to back-off or not at each stage. An exception to Stage 2 is that if a triple has no preposition, that is, \((v, Prep = e, n)\), always back-off from Stage 1 to Stage 3 because there is no preposition to back-off.

4. EXPERIMENTS

4.1 Experimental conditions

Experiments were conducted on part of the essays (Asao, 2000) written by Japanese learners of English. The essays consisted of about 7000 words, and 250 article errors were recognized by a professional rewriter of English texts.

The corpus used for estimating the conditional probabilities in the model consisted of the English concept explication in the EDR English-Japanese Bilingual dictionary (EDR, 1993) and the EDR corpus (EDR, 1993) (we will refer to the corpora as the EDR corpus, hereafter). The size of the EDR corpus amounted to about 3 million words.

In the case of triples whose head noun was plural and accompanied by "{each, every, this, that, one}", article errors were detected without the model. The reason is that it is almost certain that these triples are article errors, and thus there is no need to use the model. Triples whose head noun was modified by the possessive pronouns were not the target of the detection, because the possessive pronouns take the place of the articles. We made nine rules for the detection including the two rules above.

4.2 Experimental procedures

First, we estimated the conditional probabilities using the methods described in Section 2 and Section 3. We used "Apple Pie Parser" (Skein, 2000) to extract the triples from the EDR corpus.

Second, we set the threshold \(\theta\) in eq. (3). We randomly took 1000 sentences out of the EDR corpus, and applied eq. (3) with the nine rules mentioned in Subsection 4.1 to them. As a result of setting \(e(\theta) = 0.1\), we got \(\theta = 0.317\). We also set the cut off frequencies \(c_1 = c_2 = c_3 = 10\).

Finally, we detected article errors in the essays using the model with the parameters above. As a preprocessing, spelling errors in the essays were corrected using a spell checker.

4.3 Evaluation method

We use recall and precision to evaluate the model. \(R\) (Recall) and \(P\) (Precision) are defined as follows:

\[
R = \frac{M_{\text{correct}}}{M},
\]

\[
P = \frac{M_{\text{correct}}}{M_{\text{total}}},
\]
\[ P = \frac{M_{\text{correct}}}{M_{\text{detected}}} \]  

where \( M, M_{\text{correct}}, \) and \( M_{\text{detected}} \) are the number of article errors in the essays, the number of article errors detected correctly, and the number of detected article errors. Here, \( M = 250 \), as we have already explained.

We also define \( F \) \((F\text{-measure} \text{ (Rijisbergen, 1979)})\) as follows:

\[ F = \frac{(b^2 + 1) \cdot R \cdot P}{b^2 P + R} \]  

Here, parameter \( b \) in eq. (11) represents the relative weights of \( R \) and \( P \). We set \( b = 1 \) in the experiments.

### 4.4 Experimental results and discussion

Table 3 shows the experimental results. "Back-off" corresponds to the model with the backed-off estimate. "No back-off" corresponds to the model without the backed-off estimate. In "No back-off", the threshold was set to \( \theta = 0.334 \) using the same method as "Back-off", and triples \((v, \text{prep}, n)\) that appeared ten times or less in the EDR corpus were not the target of the detection. "Singles" corresponds to the model based on the conditional probability \( p(l_i | n) \), which is the model presented in (Nagata et al, 2004). Like "No back-off", the threshold was set to \( \theta = 0.284 \), and singles \((n)\) that appeared ten times or less in the EDR corpus were not the target of the detection.

According to Table 3, "No back-off" works worse than "Singles", contrary to our intuition. This is because the data sparseness problem often occurs; in "No back-off", the detection of about half of the examined triples was aborted because their occurrences were ten times or less in the EDR corpus. On the other hand, the aborted triples reduced to 3.5% when the back-off estimate was applied. Consequently, the performance of "Back-off" is better than that of "No back-off" and "Singles" as shown in Table 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-off</td>
<td>0.64</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>No back-off</td>
<td>0.31</td>
<td>0.70</td>
<td>0.43</td>
</tr>
<tr>
<td>Singles</td>
<td>0.58</td>
<td>0.76</td>
<td>0.66</td>
</tr>
</tbody>
</table>

We also compared the backed-off model to the model with another backed-off estimate. In this model, eq. (7) was modified to read:

Stage 2. If \( \sum_{m=1}^{6} f(l_m, v, n) > c_2 \)

\[ p(l_i | v, \text{prep}, n) = \frac{f(l_i, v, n)}{\sum_{m=1}^{6} f(l_m, v, n)} \]

and the rest of it was unchanged. Namely, the pair \((v, n)\) was used to back-off from \((v, \text{prep}, n)\) instead of \((\text{prep}, n)\) at Stage 2. Table 4 shows the results of the comparison. "Preposition" and "Verb" represent the performance of the model presented in Section 3 and that of the model above, respectively. Table 4 shows that "Preposition" works better than "Verb", which justifies the assumption we made in Section 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preposition</td>
<td>0.64</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>Verb</td>
<td>0.63</td>
<td>0.74</td>
<td>0.68</td>
</tr>
</tbody>
</table>

While it turned out that the model based on the triples with the backed-off estimate performed better than the other models, there are still about 36% of article errors (91 out of 250) that the model did not detect. Table 5 shows the major causes of the undetected article errors. "Low-count" represents low-count events. Although we used the backed-off estimate to abate the data sparseness problem, it still affected the performance of the model. Thus, it is important to reduce the data sparseness problem to improve performance. "Model" corresponds to undetected article errors that are due to the model itself. To be precise,
these undetected articles were not detected even though there were enough occurrences for estimating the conditional probabilities. This suggests that the upper bound on recall of the model with 0.317 is around 0.8. "Rule" represents the nine rules we made in Subsection 4.1. While these rules prevent false-positives, they also prevent the model from detecting article errors in some cases. "Context" represents the undetected article errors that require contextual information to be detected. Obviously, the model does not count contextual information. Therefore, the model is not capable of detecting this kind of article error. "Parsing" represents parsing errors. Article errors are not detected by the model unless head nouns are extracted by parser.

Table 5. Causes of undetected article errors

<table>
<thead>
<tr>
<th>Cause</th>
<th>Ratio [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-count</td>
<td>30</td>
</tr>
<tr>
<td>Model</td>
<td>21</td>
</tr>
<tr>
<td>Rule</td>
<td>20</td>
</tr>
<tr>
<td>Context</td>
<td>13</td>
</tr>
<tr>
<td>Parsing</td>
<td>10</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

4.5 Comparing with the rule-based method

Comparison of the experimental results with those of other methods is difficult because there is no generally accepted test set or performance baseline (Chodorow and Leacock, 2000). Besides, it would be infeasible to implement the rule-based methods precisely because not all rules used in the rule-based methods are described in their papers.

Given this limitation, we compared the presented model (the model based on the triples with the backed-off estimate) with a rule-based method (Kawai et al., 1984) whose experimental conditions are relatively similar to those of this paper. Table 6 shows the differences between the presented model and the rule-based method.

Table 6. Differences between the prepositional model and the rule-based method (Kawai et al., 1984).

<table>
<thead>
<tr>
<th>Method</th>
<th>Presented model</th>
<th>Rule-based method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target text</td>
<td>Essay</td>
<td>Technical writing</td>
</tr>
<tr>
<td>Writer</td>
<td>Japanese learners of English</td>
<td>Japanese</td>
</tr>
<tr>
<td>Limitations of input</td>
<td>No limitations</td>
<td>Excluding omission and the comparative</td>
</tr>
<tr>
<td>Target errors</td>
<td>Article errors</td>
<td>Errors concerning articles and agreement</td>
</tr>
</tbody>
</table>

Table 7 shows the results of the comparison. "Backed-off" and "Rule-based" correspond to the presented model and the rule-based method, respectively. While the rule-based method also detects errors concerning agreement, the results for "Rule-based" excludes errors concerning agreement in order to minimize the differences in the experimental conditions.

Table 7. Comparison of the presented model and the rule-based method (Kawai et al., 1984).

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-off</td>
<td>0.64</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>Rule-based</td>
<td>0.59</td>
<td>0.73</td>
<td>0.65</td>
</tr>
</tbody>
</table>

The performance of the presented model seems to be better compared to that of the rule-based method even considering the differences between the presented model and the rule-based method. Note that the implementation of the rule-based method was just for the experiments. In other words, the rule-based method will require more knowledge such as knowledge for distinguishing mass and count nouns and rules for detection for performing on other texts as good as the results here. On the other hand, similar results to Table 7 are expected for the presented model as long as essays written by similar population of learners are given because there are no tuning and limitations of input in the presented model.

Apart from the performance, the presented model has two advantages over the rule-based method. One is that rules, which are expressed as the conditional probabilities, are automatically generated in the presented...
model, while not in the rule-based method. It is a great advantage considering the cost of making rules by hand. The other is that recall and precision rates are adjustable in the presented model by threshold; the higher threshold you set, the higher recall you get. It is beneficial for error detecting systems that recall and precision rates are adjustable (Atwell, 1987).

5. CONCLUSION

In this paper, we proposed a model for detecting article errors based on the triples (v, prep, n). Experiments showed that the model performed better than the model based on just the single (n) when the backed-off estimate was applied. From the results, we conclude that the model based on the triples (v, prep, n) is effective to detecting article errors and that the backed-off estimate is crucial for the model.

We also compared the model with a rule-based method (Kawai et al., 1984) in the experiments. The performance of the model seemed to be better compared to that of the rule-based method even considering the differences in the experimental conditions. From the view of implementation, the model is superior to the rule-based method because the model is automatically constructed once a corpus is given. In addition to this, it is a good point of the model that recall and precision rates are adjustable.

There are a few possible improvements that may raise the performance of the model further. A larger size of the corpus is almost certain to improve the performance. More sophisticated method for estimating the conditional probabilities may improve the performance because about one third of undetected article errors were due to the data sparseness problem.

An article error detecting system that implements the model presented in this paper is available at: http://www.ai.info.mie-u.ac.jp/~nagata/error_detection/

ACKNOWLEDGEMENT

Our thanks go to Satoshi Sekine, who is the author of Apple Pie Parser.

REFERENCES