Natural Language Engineering **12** (2): 115–129. © 2006 Cambridge University Press
doi:10.1017/S1351324906004190 Printed in the United Kingdom

Detecting errors in English article usage by non-native speakers

N A-R A E H A N
University of Pennsylvania, 619 Williams Hall,
36th & Spruce Street, Philadelphia, PA 19104, USA
e-mail: nrh@ling.upenn.edu
and
Educational Testing Service
Rosedale Rd. MS 13E, Princeton, NJ 08541, USA

M A R T I N C H O D O R O W
Hunter College of the City University of New York
695 Park Avenue, New York, NY 10021, USA
e-mail: mchodoro@hunter.cuny.edu

C L A U D I A L E A C O C K
Pearson Knowledge Technologies, 4940 Pearl East Circle, Boulder, CO 80301, USA
e-mail: cleacock@pearsonkt.com

(*Received 1 February 2006*)

Abstract

One of the most difficult challenges faced by non-native speakers of English is mastering the system of English articles. We trained a maximum entropy classifier to select among *a/an*, *the*, or *zero* article for noun phrases (NPs), based on a set of features extracted from the local context of each. When the classifier was trained on 6 million NPs, its performance on published text was about 83% correct. We then used the classifier to detect article errors in the TOEFL essays of native speakers of Chinese, Japanese, and Russian. These writers made such errors in about one out of every eight NPs, or almost once in every three sentences. The classifier's agreement with human annotators was 85% (kappa = 0.48) when it selected among *a/an*, *the*, or *zero* article. Agreement was 89% (kappa = 0.56) when it made a binary (yes/no) decision about whether the NP should have an article. Even with these levels of overall agreement, precision and recall in error detection were only 0.52 and 0.80, respectively. However, when the classifier was allowed to skip cases where its confidence was low, precision rose to 0.90, with 0.40 recall. Additional improvements in performance may require features that reflect general knowledge to handle phenomena such as indirect prior reference. In August 2005, the classifier was deployed as a component of Educational Testing Service's Criterion^{SM} Online Writing Evaluation Service.

1 Introduction

As any teacher of English as a Second Language can attest, one of the most complex problems faced by a non-native speaker is when to use *a* (or *an*), *the*, or *0* (*zero* or *no*) article at the beginning of a noun phrase (NP). This is particularly

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

problematic for speakers of Japanese, Chinese, Korean, Russian, and other languages that do not have articles. The goal of our work is to develop tools that provide feedback to these writers and others when they choose an article (*a* instead of *the*, or vice versa), fail to use an article when one is required (**I kicked ball*), or use an article when there should be none (**I want a knowledge*). Of course, determining correct article usage is valuable for more than just second language learning. It is crucially important for high quality machine translation (MT), as well as for text summarization, text generation, and a host of other applications ranging from optical character recognition to text-to-speech devices for the disabled.[1]

In this paper, we describe the performance of a maximum entropy classifier (Ratnaparkhi 1997, 1998) for English articles that is trained on up to 8 million noun phrases (NPs) extracted from a corpus of published text. (For a brief description, see Han, Chodorow, and Leacock 2004.) The system uses local context features in the form of words and part of speech tags to compute the probability that the NP will have *a/an*, *the*, or *0* article. The system's performance is evaluated in two ways: (i) on held-out data from the same corpus as the training set, and (ii) on essays written for the Test of English as a Foreign Language (TOEFL) by native speakers of Japanese, Chinese, and Russian. Before describing in greater detail the classifier and its performance, we first discuss some of the factors that make this task so challenging and then review related prior work.

2 Selection of English articles

The distribution of articles is complex largely because it reflects the interaction of many heterogeneous factors. Some are lexical, such as the countability of the head noun in the NP, which determines if the indefinite article *a* can be used (**a knowledge*). This property of countability is not dichotomous, as reflected in Allan's (1980) noun countability preferences that range from fully countable (e.g., *boy*) at one extreme to strongly countable (*cake*), weakly countable (*beer*), and uncountable (*knowledge*) at the other extreme (see also Bond, Ogura and Ikehara 1994). Baldwin and Bond (2003) have shown that some countability properties of a noun can be learned from local syntactic contexts derived from corpus data. But knowing that a noun is countable is not enough to predict its article. This is due in part to the fact that countability is actually a property of the word sense rather than the word (compare *paper* vs. *a paper*). Countability has also been shown to be at least partly predictable from semantic class (Bond and Vatikiotis-Bateson 2002), but even for fully countable words, certain semantic classes, such as units of measure or time, generally take *a* (*the car needs a/*the gallon of gasoline*), while others, such as body parts, typically require *the* (*he was hit in the/*an eye*).

Syntactic properties of the NP are also important for article selection. When a superlative adjective like *best* or an ordinal like *first* is a premodifier of the noun, there is a strong preference for *the* (e.g., *the/*a best actor*). The presence of a

[1] See Knight and Chander (1994) and Minnen, Bond and Copestake (2000) for a discussion of these and other possible applications.

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

complement of the NP can override a lexical factor. For example, even uncountable nouns, such as *knowledge*, can take the article *a* if there is a prepositional phrase (PP) complement (e.g., *a knowledge of Spanish*). The type of phrase and the head of the phrase in which the NP is embedded may also play a role. When a measure or time noun is the object of a by PP, *the* is used instead of *a* (*paid by the/*an hour*; *sold by the/*a gallon*).

Discourse factors affect definiteness and the selection of *a* or *the* or the use of the noun in its plural form. *A* is used for the first mention of an entity in a discourse and *the* is used to refer back to an earlier mentioned entity, but, of course, there are many qualifications to these generalizations. For example, generic uses typically take *the* or plural forms even when first mentioned, and ascriptive uses (Bond, Ogura and Kawaoka, 1995) usually take *a* (*the computer is a useful tool*). Sometimes the reference to an earlier entity is only indirect (*she found an umbrella but the handle was missing*) and must be filled in from general knowledge.

General knowledge has an effect on article selection in other ways as well. Unique entities are modified by *the* (*the Moon*), as are items that are familiar to the discourse participants (*he read the morning paper*). Some rules are highly constrained and require quite specific knowledge. *A* is used if a proper noun refers to a manufacturer's product, but no article is used if it refers to the manufacturer's stock (compare *he bought an IBM* vs. *he bought IBM*). Quirk, Greenbaum, Leech and Svartvik (1985) and Pica (1983) discuss many of these and other usage rules.

Unfortunately, almost every rule for articles has many exceptions or subrules, and the interactions that occur when two or more rules apply can be very difficult to predict. Despite these complexities, some researchers have reported success in using handcrafted rules for generating articles in Japanese-to-English MT (Murata and Nagao 1993; Bond and Ikehara 1996; Bond *et al.* 1995) and Japanese-to-German MT (Heine 1998). Others have turned to machine learning and text corpora to induce a set of decision trees for selecting *a* or *the* (Knight and Chander, 1994) or a set of feature weights (Minnen, Bond and Copestake 2000) for deciding which article is required.

3 Related research

Most of the research on article selection has been carried out using handcrafted rules for translation. The goal of Murata and Nagao's (1993) work was to provide a rich enough representation of a Japanese source document to support article generation for an English translation. They described a set of 84 heuristic rules for determining the referential property of Japanese NPs (generic, definite, indefinite) and 48 heuristic rules for determining NP number (singular, plural, and uncountable). When these were applied to Japanese test materials, the results showed 69 correct labeling on the generic-definite-indefinite classification and 85.6% correct on the singular-plural-uncountable distinction. Bond *et al.* (1995) described a hierarchy of heuristics for articles and NP number that use local syntactic context in the Japanese source plus semantic information for 2,800 Japanese noun and verb categories (e.g., FOOD, BODILY ACTION). These rules were implemented in the ALT-J/E system, and they

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

improved its translation of articles and number significantly, from 65% correct to 77% correct. Bond and Ikehara (1996) described heuristics for countability and referentiality which, when incorporated into the ALT-J/E system, produced correct translations of articles and number in 80% of the NPs tested. Heine (1998) used hierarchically organized heuristic rules for assigning definiteness to Japanese NPs in about 1,000 sentences taken from an appointment scheduling domain. The rules applied to 79.5% of the NPs in the test data, and for these they were correct 98.9% of the time. When the default, "definite", was assigned to the remaining NPs, overall accuracy was about 90%.

In contrast to these handcrafted heuristics, automatic generation of rules was the focus of Knight and Chander's (1994) work. They trained a decision-tree builder on 400,000 NPs from *Wall Street Journal* text, each beginning with *a/an* or *the*.[2] For each training example, lexical features were extracted, such as the head noun, the premodifying words, and the two words following the head. More abstract features that were also used included the parts of speech of the lexical features and other subcategory information, such as the whether the head was a mass noun and whether a premodifying adjective was superlative. From these, a decision tree with 30,000 features and over 200,000 rules was automatically constructed to distinguish between NPs with *a/an* and *the*. As a lower bound for performance, Knight and Chander noted that simply guessing the more common article, *the*, would yield the correct answer in 67% of the training examples. As an upper bound, they asked human subjects to predict the article from the same context available to the system, i.e., the core NP (without the article), plus the two words that precede the NP and the two words that follow it. With this information, humans achieved 83% to 88% accuracy. To test their system, Knight and Chander built decision trees for the 1,600 most frequent head nouns in their corpus, accounting for about 77% of the NPs in their test set. On these, they achieved 81% accuracy. When the remaining NPs were blindly assigned *the*, the overall performance on the test set was reported as 78% correct.

Minnen *et al.* (2000) extracted eight different types of features from over 300,000 NPs having *a/an*, *the*, or *zero* article in the Penn Treebank *Wall Street Journal* data. The features included the head of the NP, its functional tag in the Treebank, and its part of speech tag; the category of the constituent embedding the NP and its functional tag in the Treebank (e.g., SUBJ); the presence of a determiner in the NP; and the countability preference of the head and the head's semantic class (both from the transfer dictionary of the ALT-J/E system). The researchers used the TiMBL 3.0 memory-based learner (Daelemans *et al.* 2000), which stores all the training examples in memory and classifies test examples by inducing a class from the most similar stored examples. TiMBL implements a number of different learning methods. For their article prediction task, Minnen, et al. used its modified value difference metric (MVDM) for distance and a k-nearest neighbors algorithm. In their test materials, 70% of all NPs had the *zero* article. Their system was significantly better than this baseline as it achieved an accuracy of 83.6% when all the features were combined.

[2] Cases of the *zero* article were not considered in Knight and Chander (1994).

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Using the head of the NP alone yielded about 80% accuracy. Other single features that produced above baseline performance were the semantic class of the head and the NP's functional tag.

The work of Izumi, Uchimoto, Saiga, Supnithi and Isahara (2003) and Izumi, Uchimoto and Isahara (2004) is most similar to our own. They used a maximum entropy classifier to detect 13 different types of grammatical and lexical errors made by Japanese speakers in an interview-based test of spoken English. Of these, article errors were, by far, the most common type. The classifier was trained using 5,600 error-annotated sentences from the transcribed spoken corpus, plus 105,000 correct sentences from the same source and 7,500 sentences in which artificial errors were introduced by the researchers. For the choice of features, the classifier relied on local and surface features at the target location and within a window of $\pm$ two positions. The features consisted of the words, their root forms and their parts-of-speech. Izumi *et al.* (2004) report a moderate level of success: 62% recall and 78% precision for detecting missing articles ("omission-type errors"), and 25% recall and 59% precision for detecting incorrectly used or extraneous articles ("replacement/insertion-type errors").

4 Relation to previous work

The approach we have used differs in several ways from the other machine learning systems for article selection.

1. Instead of training on one source, the *Wall Street Journal* or a corpus of spoken English, we have used a subset of text from a diverse corpus of English. The MetaMetrics, Inc. text corpus is a collection of approximately 23,000 text files, about 500 million words in all, consisting of current English fiction, non-fiction and textbooks from kindergarten to graduate school reading level. Corpus diversity poses a greater challenge for any statistical classifier as different genres of writing are likely to have different proportions of generic usage (e.g., science texts vs. short stories) and a more varied array of word senses. It is precisely for these reasons that we have chosen a multi-source dataset to build a model for student essays written on TOEFL. In this way, we hope to reduce the inevitable loss which is experienced when one type of text is used in training and another is used in testing.
2. We have trained on much larger sets than earlier studies, up to 8 million NPs, in the hopes that greater lexical coverage in training will support better performance in testing. Previous studies (Minnen *et al.* 2000; Knight and Chander 1994) have shown that the head noun is the most important feature in article selection, and that classifier performance improves with more training. The 8 million NPs we have used constitute a set that is many times larger than those of previous studies. This should provide us with a much clearer picture of the effect of training frequency on performance.
3. We have used as features only words, part-of-speech tags, positions relative to NP boundaries, and corpus-based frequency measures. In particular, we have avoided using semantic information and other features found in hand-coded

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

dictionaries. Our intent was to produce a system that would automatically adapt to its training input without the need for additional knowledge sources.

4. In an approach similar to that of Izumi *et al.* (2003, 2004), we have employed a maximum entropy model (Ratnaparkhi 1997, 1998) to estimate the probability of *a/an*, *the* and *zero* article for NPs, based on their local contextual features. Maximum entropy has been shown to perform well in combining heterogeneous forms of evidence, as in word sense disambiguation. It also has the desirable property of handling interactions among features without having to rely on the assumption of feature independence, which is quite obviously false in the case of article selection.

In summary, the research described in the next section differs from most previous efforts inasmuch as it uses a highly diverse corpus, very large training sets, non-semantic corpus-derived features, and a maximum entropy model for making three-way choices between *a/an*, *the*, and *zero* article.

5 Building a model

From the MetaMetrics corpus, a total of 721 text files, containing 31.5 million words, were selected from 10th through 12th grade reading levels. For purposes of four-fold cross-validation, the files were organized into four groups of approximately 180 each. Each file was tagged with a maximum entropy part of speech tagger (Ratnaparkhi, 1996), and it was then chunked into NPs by a maximum entropy noun phrase chunker provided by Thomas Morton (`http://opennlp.sourceforge.net/`). In total, there were about 8 million NPs, with approximately 2 million in each of the four file groups.

Following chunking, the NPs were converted into sets of features based on the local context. The local context consisted of the two words before the beginning of the NP (pre-pre-NP and pre-NP), the words within the NP (excluding, of course, the article if there was one), and the word following the NP (post-NP). There were 11 local feature types in all (see Table 1). Most of the features combined lexical and syntactic information, e.g., the head word and its part of speech tag (head/PoS). One feature, countability of the head noun, was based, not on the local context, but on the proportion of occurrences of the noun in its plural form in the corpus. If the plural accounted for less that 3% of all occurrences of the noun, then the countability value was "uncountable". If a noun occurred in plural form more than 95% of the time, then it was classified as "pluralia tantum", i.e. only plural form, as in *scissors*. Nouns in the middle range received the value "countable"; those that had not been observed in the corpus were assigned "unknown".

For training, three of the four sets of files were used. Each of the approximately 6 million NP training events consisted of the features of the NP along with the article that had occurred with it (*a/an*, *the*, or *0*). For a feature to be included, it had to occur a minimum of five times in the training events. On average, there were about 390,000 features in each of the four maximum entropy models, a number that reflects the many lexical values of the head word and other elements of the NP

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Table 1. *Accuracy of single features used in the classifier, with a default selection of* 0 *articles for unknown values (PoS = part of speech tag) (Adapted from Han* et al. *(2004))*

Feature	% Correct
word/PoS of all words in NP	80.41
word/PoS of pre-NP + head/PoS	77.98
head/PoS	77.30
PoS of all words in NP	73.96
word/PoS of post-NP	72.97
word/PoS of initial word in NP	72.53
PoS of initial word in NP	72.52
word/PoS of pre-NP	72.30
PoS of head	71.98
head's countability	71.85
word/PoS of pre-pre-NP	71.85
none: defaulting to *0* determiner	71.84

context, and no attempt was made to remove potentially extraneous features. The classifier was trained to convergence using generalized iterative scaling.

6 Test results for published text

For each cross-validation test, the features of the NPs in the held-out set of files were presented to the classifier, which computed the probabilities of the outcomes *a/an*, *the* and *0*. The classifier was scored as correct if the article that it selected as the most probable was the one that had actually occurred with the NP.

The most common article in the corpus was the *zero* article (71.84% of all NPs), followed by *the* (20.05%), and *a/an* (8.10%). Across the four cross-validation runs, performance of the classifier ranged from 87.59% to 88.29% correct, with an average of 87.99 %, well above the baseline of 71.84% that would result from always assigning the *zero* article.

The contribution of each feature was assessed by building a model with only that feature and a default that allowed the classifier to select the most common article (*0* article) when the feature value did not appear in training (see Table 1). Under these conditions, the most predictive single feature was the entire noun phrase, i.e., the concatenation of all of the words and part of speech (PoS) tags in the NP. We would expect this "whole NP" feature to work well when the corpus size is very large, as in the current study. The next best feature was the combination of the word before the beginning of the NP (pre-NP) and the head. This combination represents information about the interaction between the embedding constituent and the head word. In particular, it captures the behavior of certain nouns when they are used as objects of prepositions (cf. *a/the summary* vs. *in summary*). The head with its part of speech was the next best predictor. Somewhat surprisingly, performance using the countability feature of the head was only slightly better than the baseline. This may be due to the particular relative frequencies that were used for assigning the

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Table 2. *Accuracy as a function of training set size (adapted from Han* et al. *(2004))*

Training set size (number of NPs)	% Correct
75,000	83.03
150,000	83.49
300,000	84.92
600,000	85.75
1,200,000	86.59
2,400,000	87.27
4,800,000	87.92
6,000,000	87.99

Table 3. *Accuracy for NPs headed by nouns, as a function of frequency of the head (adapted from Han* et al. *(2004))*

Frequency of head noun	% Correct
1	73.6
5	73.6
10	76.0
50	78.5
100	79.6
500	80.7
1,000	81.9
5,000	82.4
10,000+	86.3

countability values "countable", "uncountable", and "always plural". The entries in Table 1 show the performance of all 12 features.

Another way to measure the contribution of each feature is to note the effect on performance when it is removed from the full model consisting of all the features. The resulting decrement reflects a contribution that cannot be compensated for by the other features. Removing the feature `word/PoS of pre-NP + head/PoS` produced the largest reduction in accuracy, about 1%.

Table 2 shows accuracy as a function of the size of the training set (in number of NPs). As expected, performance improved as the training sets grew larger. Minnen, et al. (2000) reported 83.6% correct when they trained their classifier on the same type of three-choice article selection problem using 300,000 NPs from the *Wall Street Journal*. With a comparable amount of training, our results are about 1.4% better on the NPs of the MetaMetrics corpus.

For NPs headed by nouns, performance improved as a function of the number of occurrences in the 31.5 million word corpus, as shown in Table 3.

Mean performance by type of article was 63.53% correct for *a/an*, 72.14% for *the*, and 95.25% for *0* article. These differences undoubtedly reflect many factors, including the syntactic type of the NP head and the referential use of the NP in discourse. With regard to syntactic type, NPs headed by plural nouns do not take *a/an*, so for these, there are only two choices, *the* or *0* article. When the head is a

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Table 4. *Accuracy for various syntactic types of NP head (adapted from Han* et al. *(2004))*

Syntactic type of head	% Correct
Singular noun	80.99
Plural noun	85.02
Personal Pronoun	99.66
Proper noun singular	90.42
Proper noun plural	82.05
Number	92.71
Demonstrative pronoun	99.70
Other	97.81

pronoun, the *0* article is almost always correct. Table 4 shows system accuracy by syntactic type of the head. As expected, the most difficult cases are NPs headed by singular nouns.

The inclusion of NPs headed by personal or demonstrative pronouns allows us to compare the overall performance of our system with previous research, but including these NPs also inflates measures of system performance since it is trivial to select the correct (*zero*) article in these cases. In order to produce a more realistic measure, we excluded all NPs not headed by singular or plural common or proper nouns (tags `NN`, `NNS`, `NNP`, and `NNPS` in the Penn tag set), and to keep the same training set size, we added more text from the corpus until the number of such noun-headed NPs reached 8 million. The zero article category, though still the most common, represented 54.40% of these NPs, as compared to 71.84% in the original training set. When the classifier was trained and tested using four-fold cross-validation, its average accuracy was 83.00%. Although overall performance was less accurate than before, it was significantly higher than the reduced baseline of 54.40%.

After the cross-validation tests were run, a maximum entropy model was built from the entire set of 8 million noun-headed NPs. It is this model that was used to detect errors in TOEFL essays.

7 Article errors in TOEFL essays

As we observed in the introduction, mastering the English articles is one of the most daunting tasks facing the non-native speaker – especially when L1 does not have articles. To document the extent of the problem, we examined 668 TOEFL essays written by native speakers of Chinese (225 essays), Japanese (234 essays), and Russian (184 essays). The essays had been written for a number of different topics and had been scored for overall quality of writing by two readers on a six-point scale ranging from 1-lowest to 6-highest.[3] 4 is the most frequent of the 6 score points, whose distribution is shown in Table 5.

[3] A detailed explanation of these scores is available on the web at: `http://www.ets.org/toefl/learners/pbt/writingscore.html`.

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Table 5. *The proportions of TOEFL essays at each score point*

Score point	% Essays
1's	1%
2's	5%
3's	28%
4's	40%
5's	15%
6's	12%

Table 6. *Proportion of text units containing one or more article errors for three language groups*

Language Group	Essay	Sentence	NP
Chinese	0.95	0.30	0.12
Japanese	0.98	0.34	0.15
Russian	0.97	0.32	0.13
Mean	0.97	0.32	0.13

For the current study, two annotators classified NPs for correct usage with these seven categories: (1) extraneous article (*a/an* or *the* was used but *0* article was correct), (2) *a-the* confusion (*a/an* instead of *the*, or vice versa), (3) missing *a/an*, (4) missing *the*, (5) missing either article (an article was missing but *a/an* or *the* would be equally correct), (6) unable to determine correct usage (a garbled sentence, a non-article error, or a context in which it was impossible to determine the intent of the writer), (7) correct usage. The annotators, who had access to the full text of the essays, independently categorized a subset of the NPs, agreeing 98% of the time on error categories 1–5, with a kappa of 0.86.

We examined their annotations for all 29,759 NPs headed by nouns. Table 6 shows the presence of article errors (collapsed over categories 1–5) for text units of various sizes by native language. Almost every essay contained at least one error, about one-third of the sentences had errors, and, on average, 13%, or about 1 in every 8 NPs had an error. These data confirm the generally held belief that difficulties in using articles are indeed quite common among native speakers of these languages.

Distributions by language group are shown in Table 7 for the six error categories including "other" (unable to determine). We were surprised by the relatively low proportion of *a-the* confusions compared to the much higher rates of omissions.

8 Test results for TOEFL essays

To compare the performance of the classifier with the annotators' judgments, we gave the classifier the NPs in the TOEFL essays and mapped its article selections onto the annotators' error categories 1–5 and category 7 "correct usage". NPs that the annotators could not judge, category 6 above, were removed from the test set

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Table 7. *Proportion of article errors by error type for three language groups*

Error Type	Chinese	Japanese	Russian
1. Extraneous	0.238	0.215	0.222
2. *a-the* confusion	0.064	0.062	0.050
3. Missing *a/an*	0.291	0.334	0.315
4. Missing *the*	0.185	0.223	0.261
5. Missing either	0.025	0.029	0.018
6. Other	0.198	0.136	0.134
Total	1.000	1.000	1.000

(about 16% of the total). For category 5, "missing either article", the classifier was scored as in agreement with the annotator if it selected either *a/an* or *the*.

One major difference between the annotators and the classifier was in the treatment of misspelled words. The humans generally rendered a judgment based on their knowledge of the correct form of the word. The classifier, however, treated the misspelled form as unknown. In an interactive writing analysis system, the misspelled word would be highlighted and the writer would be asked to correct it, but in our batch-testing mode, this option was not available, so we removed NPs whose head nouns were misspelled (about 4.5% of the total). This left 22,322 NPs containing 2,703 errors.

Results of the test showed 85% agreement (kappa = 0.48) between the classifier and the human. The kappa value, which indicates only a fair level of agreement, reflects the high proportion of times the category "correct" was used by the human annotators and by the system. With so many "correct" annotations, the baseline of agreement expected by chance was about 72%.

An analysis of the classifier's judgments revealed that 79% of the errors in the test set were correctly detected, but the classifier also produced many false positives, resulting in only 44% precision. We wondered if this rather poor performance was related to the difficulty of making a three-way decision about article usage, so we retrained the classifier on a simpler binary choice for each NP – `yes`, it should have an article, or `no`, it should not have an article. Results of the test showed 89% agreement (kappa = 0.56) between the binary classifier and the human. With the binary model, recall for the test set was 80%, about the same as the three-way model, and there was a modest improvement in precision, to 52%.

In a large proportion of the classifier's false positives, the category it chose was just slightly higher in probability than its second choice. For example, one essay contained the sentence *The books are assigned by professors*. The annotators marked *professors* as correct whereas the classifier selected *the*, the choice for which it had computed a very slightly higher probability (0.51) than for the *0* article (0.49). Both choices are grammatical in terms of the local context. In cases such as this, it should be easy for the system to recognize that its selection is not clear-cut, and therefore that it should not decide between the two alternatives. We re-ran the binary classifier, this time allowing it to skip an NP if the difference between the probability of its first and second choices was less than 0.75, which we set as the threshold or cutoff

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Table 8. *Precision and recall of the binary classifier with a decision threshold of 0.75*

Error Type	Precision	Recall
1.Extraneous article	0.80	0.34
3,4,5 Missing article	0.94	0.42
Combined	0.90	0.40

Table 9. *Precision and recall of the three-way classifier with a decision threshold of 0.75*

Error Type	Precision	Recall
1. Extraneous article	0.79	0.39
3. Missing *a*	0.95	0.20
4. Missing *the*	0.89	0.22
5. Missing either	0.90	0.20

for making a decision. Under these conditions, precision rose to 90%, although recall dropped to 40%. In terms of providing information to second language learners, we chose to trade recall for precision in the belief that it is better to miss some errors than to mark correct article usage as wrong. The results for the modified binary classifier are shown in Table 8, where categories 3, 4 and 5 have been combined, since the classifier's output does not specify which article is missing. Precision and recall for missing articles were substantially better than for extraneous ones.

The decision threshold of 0.75 was set empirically to give 90% overall precision for the test set. We wondered if this value would generalize to a new evaluation set of TOEFL essays, so we drew another text sample consisting of 4,200 noun-headed NPs, annotated them and ran them through the binary classifier. The results were quite similar to those of the test set. For extraneous articles, precision and recall were 0.78 and 0.36, respectively; for missing articles, they were 0.93 and 0.41, respectively.

The three-way classifier did not fare as well on the test set as the binary classifier had. Its overall performance was about the same for extraneous articles, but its recall was only about half that of the binary classifier for missing articles, categories 3–5, as shown in Table 9.

The binary classifier found about twice as many missing article errors as the three-way classifier, but the latter provided more information about each error because it indicated which article was needed to correct the error. To take advantage of the strengths of both systems, we combined them: When the binary classifier identified a missing article, the three-way classifier was then used to determine which one was appropriate. In this way, the combined system was able to detect 42% of missing articles with 94% precision, and in more than half of these cases, it also indicated whether *a/an* or *the* was needed to fix the error.

Finally, we examined the three-way classifier's performance on *a-the* confusions, an error type which accounted for only about 6% of the total number of errors in

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

the test set (see Table 6). Using 0.75 as the decision threshold, performance was quite poor (0.33 precision and 0.17 recall). Even with the threshold set at 0.95, precision was just 0.67, while recall was only 0.11.

Why did the system perform so poorly on *a-the* confusions? One obvious reason might be the classifier's lack of discourse information. As noted earlier, *a* is generally used for the first mention of an entity in a discourse and *the* is used when referring back to it. Accordingly, we added to the model a feature that indicated whether the head noun had occurred among the previous twenty-five NPs in the essay and, if so, what type(s) of article it had on those occasions. Much to our surprise, the addition of this information about prior occurrences in the discourse had no significant effect on system performance. When we examined the false positives, the NPs that the system identified as wrong but which human annotators considered to be correct, some of the reasons became clear. Many head nouns first appeared with *the*, as in *A student will not learn if she hates the teacher*. The system selected *a* for *teacher*. Here, arguably, we have indirect prior reference, comparable to our earlier example of *an umbrella with the handle missing*. It may be possible to identify these cases automatically based on semantic relations (e.g., part/whole, synonymy) or from co-occurrence data (the statistical association between *student* and *teacher*). In other instances, not only the first appearance but also all subsequent appearances of a head noun took the indefinite article. For example, a student wrote about his desire to obtain a scholarship and the possibilities that *a scholarship* would afford. This may be a case of opaque context (Gamut 1991), where *a scholarship* does not refer to any particular scholarship but is instead a description or sense of what the student wants. Opaque contexts can be supported by verbs such as *want*, *desire*, and *look for*.

More than two-thirds of the false positives for the extraneous article category were cases in which the writer had used *the* with a plural noun. For example, the system marked *Students should choose the classes they want* as an error, preferring the equally grammatical *Students should choose classes they want*. A plural without an article often indicates a generic usage, but the distinction can be quite subtle, as in this example. Our decision threshold was intended to allow the classifier to skip such cases, but the system's confidence was high for its judgment that *classes* should not have a definite article. In part, this may be due to a mismatch between the proportions of particular plural nouns with *the* in the training corpus and in our student essay test set.

In about 85% of false positives, combined across all error categories, the system's suggestion was, in fact, grammatical, as in the cases described above. When the suggestion was ungrammatical, it was usually due to tagging or NP chunking errors, difficulties with conjoined head nouns, or long distance phenomena that were beyond the range of the local context. An example of a long range dependency is *The more people attend the concerts, the more security we will need*. Here, a template for *the more … the more* constructions might help prevent false positives.

In August 2005, the combined binary + three-way classifier was added to the set of writing analysis tools used by Educational Testing Service's Criterion^{SM} Online Writing Evaluation Service (Burstein, Chodorow and Leacock 2003). Criterion is a

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

web-based application that evaluates students' essays and provides a holistic score indicating overall writing quality. It also highlights specific errors in grammar, usage, mechanics, and style, and provides diagnostic information along with suggested corrections. Currently, there are more than 900,000 users of Criterion, many of them non-native speakers of English. We look forward to their feedback and to the valuable information that a community of users can provide to help us improve our system's performance in detecting and correcting article errors.

9 Conclusion

The combination of a maximum entropy classifier and a very large training corpus of heterogeneous documents has yielded results that are better than those previously reported. The main advantage of this approach is that it is fully automated and does not require additional lexical or knowledge resources. Some of its remaining deficiencies are in its handling of generic usage, indirect reference, and opaque contexts. Despite this, we believe that a system which detects a large proportion of the non-native writer's errors involving articles will prove to be a valuable tool for language instruction and for language assessment.

Acknowledgments

We wish to thank Tom Morton for his development of our maximum entropy part of speech tagger and NP chunker, and for his help in setting up the classifier. We are also grateful to Shauna Cooper, Todd Farley, and Irma Lorenz for annotating errors in the TOEFL essays, to Eleanor Bolge for developing the error annotation GUI, and to Derrick Higgins and Neil Dorans for helpful comments on an earlier version of this paper.

References

Allan, K. (1980) Nouns and countability. *Language, 56*, 541–567.

Bond, F. and Ikehara, S. (1996) When and how to disambiguate? Countability in machine translation. *International Seminar on Multimodal Interactive Disambiguation: MIDDIM-96*, pp. 149–160. Grenoble, France.

Bond, F., Ogura, K. and Ikehara, S. (1994) Countability and number in Japanese to English machine translation. *Proceedings of Coling '94*, pp. 32–38.

Baldwin, T. and Bond, F. (2003) Learning the countability of English nouns from corpus data. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Bond, F., Ogura, K. and Kawaoka, T. (1995) Noun phrase reference in Japanese-to-English machine translation. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '95)*, pp. 1–14.

Bond, F. and Vatikiotis-Bateson, C. (2002) Using an ontology to determine English countability. *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

Burstein, J., Chodorow, M. and Leacock, C. (2003) Criterion: Online essay evaluation: an application for automated evaluation of student essays. *Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico.

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A. (2000) *TiMBL: Tilburg Memory Based Learner, version 3.0, Reference Guide*. ILK Technical Report 0001.

Gamut, L. T. F. (1991) *Logic, Language, and Meaning*, Chicago: University of Chicago Press.

Han, N.-R., Chodorow, M. and Leacock, C. (2004) Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Heine, J. (1998) Definiteness predictions for Japanese noun phrases. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98*, pp. 519–525. Montreal, Canada.

Izumi, F., Uchimoto, K., Saiga, T., Supnithi, T. and Isahara, H. (2003) Automatic error detection in the Japanese learners' English spoken data. *ACL-2003 Interactive Posters and Demonstrations: Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Izumi, F., Uchimoto, K., and Isahara, H. (2004) SST speech corpus of Japanese learners 'English and automatic detection of learners' errors. *ICAME (International Computer Archive of Modern and Medieval English) Journal*, No. 28, 31–48, Bergen, Norway.

Knight, K. and Chander, I. (1994) Automated postediting of documents. *Proceedings of the Twelfth National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA.

Minnen, G., Bond, F. and Copestake, A. (2000) Memory-based learning for article generation. *Proceedings of the 4th Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, pp. 43–48. New Brunswick, NJ.

Murata, M. and Nagao, M. (1993) Determination of referential property and number of nouns in Japanese sentences for machine translation into English. *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 218–225.

Pica, T. (1983) The article in American English: What the textbooks don't tell us. In: N. Wolfson and E. Judd (eds.), *Sociolinguistics and Language Acquisition*, pp. 222–233. Rowley, MA: Newbury House.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. Longman.

Ratnaparkhi, A. (1996) A maximum entropy model for part-of-speech tagging. *Proceedings of the First Empirical Methods in Natural Language Processing Conference*, pp. 133–141. Philadelphia, USA.

Ratnaparkhi, A. (1997) A simple introduction to maximum entropy models for natural language processing. IRCS Technical Report 97-08, Institute for Research in Cognitive Science, Philadelphia, USA.

Ratnaparkhi, A. (1998) *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD Dissertation, University of Pennsylvania.

https://doi.org/10.1017/S1351324906004190 Published online by Cambridge University Press