Detecting Errors in English Article Usage with a Maximum Entropy Classifier Trained on a Large, Diverse Corpus

Na-Rae Han
University of Pennsylvania
619 Williams Hall
36th & Spruce Street
Philadelphia, PA 19104
nrh@ling.upenn.edu

Martin Chodorow
Hunter College of the City University of New York
695 Park Avenue
New York, NY 10021
mchodoro@hunter.cuny.edu

Claudia Leacock
Educational Testing Service
Rosedale Road, 18E
Princeton, NJ 08541
cleacock@ets.org

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, based on a set of features extracted from the local context of each. When the classifier was trained on 6 million noun phrases, its performance was correct about 88% of the time. We also used the classifier to detect article errors in the TOEFL essays of native speakers of Chinese, Japanese, and Russian. Agreement with human annotators was about 88% (kappa = 0.36). Many of the disagreements were due to the classifier's lack of discourse information. Performance rose to 94% agreement (kappa = 0.47) when the system accepted noun phrases as correct in cases where its own confidence was low.

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. This is particularly 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. (See Knight & Chander (1994) and Minnen, Bond, & Copestake (2000) for a discussion of these and other possible applications.)

In this paper, we describe the performance of a maximum entropy classifier (Ratnaparkhi, 1997) for English articles that is trained on up to 6 million noun phrases (NPs) extracted from a corpus of published text. 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, and (ii) on essays written for the Test of English as a Foreign Language (TOEFL) by native speakers of Japanese, Chinese, and Russian.

2 Related Research

Most of the early research on article selection has been carried out using handcrafted rules to improve quality of machine translation systems. Murata and Nagao (1993), Bond, et al. (1995), Bond and Ikehara (1996), Heine (1998) are all such works applied to machine translation systems between Japanese and English.

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*. 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. 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*. (Cases of the zero article were not considered.) Knight and Chander built decision trees for the 1600 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 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. The researchers used the TiMBL 3.0 (Daelemans, et al., 2000) memory-based learner to train and test their model. 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.

3 Relation to Previous Work

The approach we have used differs in several ways from other machine learning systems for article selection: (1) Instead of training on one source, the *Wall Street Journal*,

we have used 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 (cf. 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. (2) We have trained on much larger sets than earlier studies, up to 6 million NPs, in the hopes that greater lexical coverage in training will support better performance in testing. Previous studies (Minnen, et al., 2000; Knight & 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 6 million NPs we have used constitute a set that is fifteen times larger than those of previous studies. (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 dictionaries. Our intent was to produce a system that would automatically adapt to its training input without the need for additional knowledge sources. (4) We have employed a maximum entropy model (Ratnaparkhi, 1997) 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. 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.

4 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. Each file was tagged with a maximum entropy part of speech tagger (Ratnaparkhi, 1996), and it was then chunked into NPs by a heuristic NP-chunking program. In total, there were about 8 million NPs, with were divided into 4 groups of approximately 2 million, for a 4-fold cross-validation.

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 combined lexical and syntactic information, e.g., the head word and its part of speech tag (head/PoS).

For training, 3 of the 4 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*). On average, there were about 390,000 features in the maximum entropy model, a number that reflects the many lexical values of the head word and other elements of the NP context.

5 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. Under these conditions, the most predictive single feature (see Table 1) was the entire noun phrase (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.

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

Table 1: Accuracy of single features used in the classifier, with a default selection of *0* article for unknown values. PoS = part of speech tag

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.

Training set size (number of NPs)	% Correct
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 2: Accuracy as a function of training set size

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 Table3.

Frequency of head noun	% Correct
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

Table3: Accuracy for NPs headed by nouns, as a function of frequency of the head

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 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.

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

Table 4: Accuracy for various syntactic types of NP head

6 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. To document the extent of the problem, we examined 150 TOEFL essays written by native speakers of Chinese (52 essays), Japanese (54 essays), and Russian (44 essays). In all, these essays contained 10,494 NPs. Two human annotators classified each NP for correct usage with these five 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), and (6) correct usage. The annotators, who had access to the full text of each essay, were in agreement on about 98% of the classifications, with a kappa equal to 0.86.

Distributions of the error categories by language groups are shown in Table 5. We were surprised by the relative low proportion of *a-the* confusions compared to the much higher rate of omissions. On average, these TOEFL essay writers produced one article error for every 16 NPs, or about once every 3 or 4 sentences. This confirms our belief that article problems are indeed quite common.

Error Type	Chinese	Japanese	Russian
Extraneous	0.011	0.015	0.018
a-the confusion	0.003	0.011	0.005
Missing a/an	0.014	0.024	0.019
Missing the	0.014	0.025	0.022
Missing either	0.003	0.005	0.003
Total	**0.045**	**0.080**	**0.067**

Table 5: Proportions of article errors by error type for three language groups

7 Test Results for TOEFL Essays

For the purpose of applying our model to TOEFL essay data, we re-trained our maximum entropy classifier only on NPs with a common-noun head, either singular or plural. This resulted in removing such trivial cases as NPs made of pronouns and demonstrative pronouns, which took up a significant portion of zero pronoun cases. As a result, the percentage of the zero pronoun cases was lowered to 54.40% of all cases, which also lead to the lower baseline performance. Still, the same training size of 6-million NPs were achieved by simply exploring and adding more NPs from new previously unused sections of the MetaMetrics corpus. When tested on held-out portion of data from the corpus, 83.00% of average accuracy was achieved. As expected, it is lower than the 87.99% of the previous model inclusive of all NPs, but still a very impressive performance considering the significant drop in baseline performance to 54.40% from previous 71.84%.

To compare the performance of the classifier with the annotators' performance on error detection, we gave the classifier the NPs in the TOEFL essays and mapped its article selections onto the six categories that the human annotators had used. For the category "missing either article", the classifier was scored as in agreement with the annotator if it selected either *a/an* or *the*. The following are actual examples of misused determiners correctly identified by both human annotators and the classifier.

#C.00539966.4.13 Above all, I think it is good for students to share <u>room</u> with others.

- Human: missing *a* or *an*
- Classifier: **0.841 *a/an***; 0.143 *the*; 0.014 *zero*

#J.00331471.6.11 Those excellent hitters began practicing <u>the baseball</u> when they were children, and dedicated a lot of time to become highly qualified.

- Human: superfluous determiner
- Classifier: 0.103 *a/an*; 0.016 *the*; **0.879 *zero***

The results showed 88% agreement (kappa = 0.37) between the classifier and annotator 1, and 89% agreement (kappa = 0.36) between the classifier and annotator 2. The kappa values, which indicate only a fair level of agreement, reflect the high proportion of "correct" category use by the annotators and by the system. With so many "correct" responses, chance could account for much of the observed agreement.

When we looked at the differences between the classifier and human judgments, one source of disagreement became clear almost immediately. For a large proportion of the system's decisions, the highest probability outcome was only marginally greater than the second highest. These were often cases where both the first and second choice were grammatical in terms of the local context, but the discourse required *a* because an entity was being mentioned for the first time or *the* because it was a subsequent reference. Although the classifier could not recognize these cases without discourse information, it was possible for the system to determine when the classifier's confidence was low and use that information in its decision. We re-ran the test, this time accepting NPs as correct (i.e., not suggesting a change) when the classifier's confidence was low, which we defined as a maximum outcome probability of 0.70 or less. The new results showed 94% agreement (kappa = 0.47) between the system and annotator 1, and 95% agreement (kappa = 0.47) between the system and annotator 2. The higher kappa values indicate a moderate level of agreement.

False positive errors are NPs that the system identified as wrong but which human annotators considered to be correct. In 62% of these, both the original essay and the system's suggestion were, in fact, grammatical. For example, one essay contained the sentence "Students can choose *the* courses they are interested in." Both annotators marked the NP as correct whereas the system selected the *0* article for *courses*. In either case, the result is acceptable. Another 22% of the system's false positives were due to spelling or grammatical errors in the NP or in its local context. The classifier relies heavily on words and tags as features, so it is especially sensitive to misspelling and other writing errors. We would expect a deployed application of the system to benefit greatly from automated spell-checking. About 6% of the false positives correctly identified article errors that both annotators had missed. Finally, in the remaining 10% of the false positives, the system introduced a determiner error into a well-formed NP.

8 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. Its main deficiency is a lack of information about previously mentioned entities. Adding discourse features should improve performance, but there will still be many subtleties of article usage that are beyond the classifier's capabilities. Despite this, we believe that a system which detects most of a 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 for his help in setting up the classifier. We are also grateful to Todd Farley for annotating errors in the TOEFL essays.

References

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

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

Bond, F., Ogura, K. and Kawaoka, T. (1995). Noun phrase reference in Japanese-to-English machine translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI '95), 1-14. (cmp-lg/9601008).

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

Knight, K., and Chander, I. (1994). Automated postediting of documents. In *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. In *Proceedings of the 4th Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, 43-48, New Brunswick, NJ: ACL.

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

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