Human Evaluation of Article and Noun Number Usage: Influences of Context and Construction Variability

John Lee
Spoken Language Systems
MIT CSAIL
Cambridge, MA 02139, USA
jsylee@csail.mit.edu

Joel Tetreault
Educational Testing Service
Princeton, NJ 08541
jtetreault@ets.org

Martin Chodorow
Hunter College of CUNY
New York, NY 10021
martin.chodorow@hunter.cuny.edu

Abstract

Evaluating systems that correct errors in non-native writing is difficult because of the possibility of multiple correct answers and the variability in human agreement. This paper seeks to improve the best practice of such evaluation by analyzing the frequency of multiple correct answers and identifying factors that influence agreement levels in judging the usage of articles and noun number.

1 Introduction

In recent years, systems have been developed with the long-term goal of detecting, in the writing of non-native speakers, usage errors involving articles, prepositions and noun number (Knight and Chander, 1994; Minnen et al., 2000; Lee, 2004; Han et al., 2005; Peng and Araki, 2005; Brockett et al., 2006; Turner and Charniak, 2007). These systems should, ideally, be evaluated on a corpus of learners' writing, annotated with acceptable corrections. However, since such corpora are expensive to compile, many researchers have instead resorted to measuring the accuracy of predicting what a native writer originally wrote in well-formed text. This type of evaluation effectively makes the assumption that there is one correct form of native usage per context, which may not always be the case.

Two studies have already challenged this "single correct construction" assumption by comparing the output of a system to the original text. In (Tetreault and Chodorow, 2008), two human judges were presented with 200 sentences and, for each sentence, they were asked to select which preposition (either the writer's preposition, or the system's) better fits the context. In 28% of the cases where the writer and the system differed, the human raters found the system's prediction to be equal to or better than the writer's original preposition. (Lee and Seneff, 2006) found similar results on the sentence level in a task that evaluated many different parts of speech.

Percentage	Article	Number	Example
42.5%	null	singular	*stone*
22.7%	*the*	singular	*the stone*
17.6%	null	plural	*stones*
11.4%	*a/an*	singular	*a stone*
5.7%	*the*	plural	*the stones*

Table 1: Distribution of the five article-number *constructions* of head nouns, based on 8 million examples extracted from the MetaMetrics Lexile Corpus. The various constructions are illustrated with the noun "*stone*".

2 Research Questions

It is clear that using what the author wrote as the gold standard can underestimate the system's performance, and that multiple correct answers should be annotated. Using this annotation scheme, however, raises two questions that have not yet been thoroughly researched: (1) what is the human agreement level on such annotation? (2) what factors might influence the agreement level? In this paper, we consider two factors: the *context* of a word, and the variability of its usage.

In the two studies cited above, the human judges were shown only the target sentence and did not take into account any constraint on the choice of word that might be imposed by the larger context. For PP attachment, human performance improves when given more context (Ratnaparkhi et al., 1994). For other linguistic phenomena, such as article/number selection for nouns, a larger context window of at least several sentences may be required, even though some automatic methods for exploiting context have not been shown to boost performance (Han et al., 2005).

The second factor, variability of usage, may be

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 60–63,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

Three years ago John Small, a sheep farmer in the Mendip Hills, read an editorial in his local newspaper which claimed that foxes never killed lambs. **He drove down to the paper's office and presented [?], killed the night before, to the editor.**

	NO-CONTEXT	IN-CONTEXT
lamb:	no	no
a lamb:	yes	yes*
the lamb:	yes	no
lambs:	yes	yes
the lambs:	yes	no

Table 2: An example of a completed annotation item.

expressed as the entropy of the distribution of the word's constructions. Table 1 shows the overall distribution of five article/number constructions for head nouns, i.e. all permissible combinations of number (*singular* or *plural*), and article ("*a/an*", "*the*", or the "*null article*"). A high entropy noun such as "*stone*" can appear freely in all of these, **either as a count noun or a non-count noun**. This contrasts with a low entropy noun such as "*pollution*" which is mostly limited to two construction types ("*pollution*" and "*the pollution*").

In this paper, we analyze the effects of varying context and noun entropy on human judgments of the acceptability of article-number constructions. As a result of this study, we hope to advance the best practice in annotation for evaluating error detection systems. §3 describes our annotation task. In §4, we test the "single correct construction" assumption for article and noun number. In §5, we investigate to what extent context and entropy constrain the range of acceptable constructions and influence the level of human agreement.

3 Annotation Design

3.1 Annotation Scheme

Two native speakers of English participated in an annotation exercise, which took place in two stages: NO-CONTEXT and IN-CONTEXT. Both stages used a common set of sentences, each containing one noun to be annotated. That noun was replaced by the symbol [?], and the five possible *constructions*, as listed in Table 1, were displayed below the sentence to be judged.

In the NO-CONTEXT stage, only the sentence in question and the five candidate constructions (i.e., the bolded parts in Table 2) were shown to the raters. They were asked to consider each of the five constructions, and to select yes if it would yield a good sentence *in some context*, and no otherwise[1].

	null	*a*	*the*	
			anaphoric	not anaphoric
singular	2	2	2	2
plural	2	n/a	2	2

Table 3: For each noun, two sentences were selected from each configuration of number, article and anaphor.

The IN-CONTEXT stage began after a few days' break. The raters were presented with the same sentences, but including the context, which consisted of the five preceding sentences, some of which are shown in Table 2. The raters were again asked to select yes if the choice would yield a good sentence *given the context*, and no otherwise. Among the yes constructions, they were asked to mark with an asterisk (yes*) the construction(s) most likely to have been used in the original text.

3.2 Annotation Example

In Table 2, "*lambs*" are mentioned in the context, but only in the generic sense. Therefore, the [?] in the sentence must be indefinite, resulting in yes for both "*a lamb*" and "*lambs*". Of these two constructions, the singular was judged more likely to have been the writer's choice.

If the context is removed, then the [?] in the sentence could be anaphoric, and so "*the lamb*" and "*the lambs*" are also possible. Finally, regardless of context, the null singular "*lamb*" is not acceptable.

3.3 Item Selection

All items were drawn from the Grade 10 material in the 2.5M-sentence MetaMetrics Lexile corpus. To avoid artificially inflating the agreement level, we excluded noun phrases whose article or number can be predicted with very high confidence, such as proper nouns, pronouns and non-count nouns. Noun phrases with certain words, such as non-article determiners (e.g., *this car*), possessive pronouns (e.g., *his car*), cardinal numbers (e.g., *one car*) or quantifiers (e.g., *some cars*), also fall into this category. Most of these preclude the articles *a* and *the*.

[1] Originally, a third response category was offered to the rater to mark constructions that fell in a grey area between yes and no. This category was merged with yes.

Rater	No-Context		In-Context	
	yes	no	yes	no
R1	62.4%	37.6%	29.3%	70.7%
R2	51.8%	48.2%	39.2%	60.8%

Table 4: Breakdown of the annotations by rater and by stage. See §4 for a discussion.

	No-Context		In-Context	
R1:↓ R2:→	yes	no	yes	no
yes	846	302	462	77
no	108	584	260	1041

Table 5: The confusion tables of the two raters for the two stages.

Once these easy cases were filtered out, the head nouns in the corpus were divided into five sets according to their dominant construction. Each set was then ranked according to the entropy of the distribution of their constructions. *Low entropy* typically means that there is one particular construction whose frequency dwarfs the others', such as the singular definite for "*sun*". *High entropy* means that the five constructions are more evenly represented in the corpus; these are mostly generic objects that can be definite or indefinite, singular or plural, such as "*stone*". For each of the five constructions, the three nouns with the highest entropies, and three with the lowest, were selected. This yielded a total of 15 "high-entropy" and 15 "low-entropy" nouns.

For each noun, 14 sentences were drawn according to the breakdown in Table 3, ensuring a balanced representation of the article and number used in the original text, and the presence of anaphoric references[2]. A total of 368 items[3] were generated.

4 Multiple Correct Constructions

We first establish the reliability of the annotation by measuring agreement with the original text, then show how and when multiple correct constructions can arise. All results in this section are from the In-Context stage.

Since the items were drawn from well-formed text, each noun's original construction should be marked `yes`. The two raters assigned `yes` to the original construction 80% and 95% of the time, respectively. These can be viewed as the upper bound of system performance if we assume there can be only one correct construction. A stricter ceiling can be obtained by considering how often the `yes*` constructions overlap with the original one[4]. The `yes*` items overlapped with the original 72% and 83% of the time, respectively. These relatively high figures serve as evidence of the quality of the annotation.

Both raters frequently found more than one valid construction — 18% of the time if only considering `yes*`, and 49% if considering both `yes` and `yes*`. The implication for automatic system evaluation is that one could potentially underestimate a system's performance by as much as 18%, if not more. For both raters, the most frequent combinations of `yes*` constructions were {*null-plural*,*the-plural*}, {*a-singular*,*the-singular*}, {*a-singular*,*null-plural*}, and {*the-singular*,*the-plural*}. From the standpoint of designing a grammar-checking system, a system should be less confident in proposing change from one construction to another within the same construction pair.

5 Sources of Variation in Agreement

It is unavoidable for agreement levels to be affected by how accepting or imaginative the individual raters are. In the No-Context stage, Rater 1 awarded more `yes`'s than Rater 2, perhaps attributable to her ability to imagine suitable contexts for some of the less likely constructions. In the In-Context stage, Rater 1 used `yes` more sparingly than Rater 2. This reflects their different judgments on where to draw the line among constructions in the grey area between acceptable and unacceptable.

We have identified, however, two other factors that led to variations in the agreement level: the amount of context available, and the distribution of the noun itself in the English language. Careful consideration of these factors should lead to better agreement.

Availability of Context As shown in Table 4, for both raters, the context sharply reduced the number of correct constructions. The confusion tables

[2] For practical reasons, we have restricted the study of context to *direct anaphoric references*, i.e., where the same head noun has already occurred in the context.

[3] In theory, there should be 420 items, but some of the configurations in Table 3 are missing for certain nouns, mostly the low-entropy ones.

[4] Both raters assigned `yes*` to an average of 1.2 constructions per item.

for the two raters are shown in Table 5. For the NO-CONTEXT stage, they agreed 78% of the time and the kappa statistic was 0.55. When context is provided, human judgment can be expected to increase. Indeed, for the IN-CONTEXT stage, agreement rose to 82% and kappa to 0.60[5].

Another kind of context — previous mention of the noun — also increases agreement. Among nouns originally constructed with "*the*", the kappa statistics for those with direct anaphora was 0.63, but only 0.52 for those without[6].

Most previous research on article-number prediction has only used features extracted from the target sentence. These results suggest that using features from a wider context should improve performance.

Noun Construction Entropy For the low-entropy nouns, we found a marked difference in human agreement among the constructions depending on their frequencies. For the most frequent construction in a noun's distribution, the kappa was 0.78; for the four remaining constructions, which are much more rare, the kappa was only 0.52[7]. They probably constitute "border-line" cases for which the line between yes and no was often hard to draw, leading to the lower kappa.

Entropy can thus serve as an additional factor when a system decides whether or not to mark a usage as an error. For low-entropy nouns, the system should be more confident of predicting a frequent construction, but more wary of suggesting the other constructions.

6 Conclusions & Future Work

We conducted a human annotation exercise on article and noun number usage, making two observations that can help improve the evaluation procedure for this task. First, although the context substantially reduces the range of acceptable answers, there are still often multiple acceptable answers given a context; second, the level of human agreement is influenced by the availability of the context and the distribution of the noun's constructions.

These observations should help improve not only the evaluation procedure but also the design of error correction systems for articles and noun number. Entropy, for example, can be incorporated into the estimation of a system's confidence in its prediction. More sophisticated contextual features, beyond simply noting that a noun has been previously mentioned (Han et al., 2005; Lee, 2004), can also potentially reduce uncertainty and improve system performance.

Acknowledgments

We thank the two annotators, Sarah Ohls and Waverely VanWinkle.

References

C. Brockett, W. Dolan, and M. Gamon. 2006. Correcting ESL Errors using Phrasal SMT Techniques. *Proc. ACL*.

N.-R. Han, M. Chodorow, and C. Leacock. 2005. Detecting Errors in English Article Usage by Non-Native Speakers. *Natural Language Engineering*, 1(1):1–15.

K. Knight and I. Chander. 1994. Automated Postediting of Documents. *Proc. AAAI*.

J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33:159–174.

J. Lee. 2004. Automatic Article Restoration. *Proc. HLT-NAACL Student Research Workshop*.

J. Lee and S. Seneff. 2006. Automatic Grammar Correction for Second-Language Learners. *Proc. Interspeech*.

G. Minnen, F. Bond, and A. Copestake. 2000. Memory-based Learning for Article Generation. *Proc. CoNLL/LLL*.

J. Peng and K. Araki. 2005. Correction of Article Errors in Machine Translation Using Web-based Model. *Proc. IEEE NLP-KE*.

A. Ratnaparkhi, J. Reynar, and S. Roukos. 1994. A Maximum Entropy Model for Prepositional Phrase Attachment. *Proc. ARPA Workshop on Human Language Technology*.

J. Tetreault and M. Chodorow. 2008. Native Judgments of Non-Native Usage. *Proc. COLING Workshop on Human Judgements in Computational Linguistics*.

J. Turner and E. Charniak. 2007. Language Modeling for Determiner Selection. *Proc. HLT-NAACL*.

[5] This kappa value is on the boundary between "moderate" and "substantial" agreement on the scale proposed in (Landis and Koch, 1977). The difference between the kappa values for the NO-CONTEXT and IN-CONTEXT stages approaches statistical significance, $z = 1.71$, $p < 0.10$.

[6] The difference between these kappa values is statistically significant, $z = 2.06$, $p < 0.05$.

[7] The two kappa values are significantly different, $z = 4.35$, $p < 0.001$.