Rethinking Grammatical Error Annotation and Evaluation with the Amazon Mechanical Turk

Joel R. Tetreault
Educational Testing Service
Princeton, NJ, 08540, USA
JTetreault@ets.org

Elena Filatova
Fordham University
Bronx, NY, 10458, USA
filatova@fordham.edu

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

Abstract

In this paper we present results from two pilot studies which show that using the Amazon Mechanical Turk for preposition error annotation is as effective as using trained raters, but at a fraction of the time and cost. Based on these results, we propose a new evaluation method which makes it feasible to compare two error detection systems tested on different learner data sets.

1 Introduction

The last few years have seen an explosion in the development of NLP tools to detect and correct errors made by learners of English as a Second Language (ESL). While there has been considerable emphasis placed on the system development aspect of the field, with researchers tackling some of the toughest ESL errors such as those involving articles (Han et al., 2006) and prepositions (Gamon et al., 2008), (Felice and Pullman, 2009), there has been a woeful lack of attention paid to developing best practices for annotation and evaluation.

Annotation in the field of ESL error detection has typically relied on just one trained rater, and that rater's judgments then become the gold standard for evaluating a system. So it is very rare that inter-rater reliability is reported, although, in other NLP subfields, reporting reliability is the norm. Time and cost are probably the two most important reasons why past work has relied on only one rater because using multiple annotators on the same ESL texts would obviously increase both considerably. This is especially problematic for this field of research since some ESL errors, such as preposition usage, occur at error rates as low as 10%. This means that to collect a corpus of 1,000 preposition errors, an annotator would have to check over 10,000 prepositions.[1]

(Tetreault and Chodorow, 2008b) challenged the view that using one rater is adequate by showing that preposition usage errors actually do not have high inter-annotator reliability. For example, trained raters typically annotate preposition errors with a kappa around 0.60. This low rater reliability has repercussions for system evaluation: Their experiments showed that system precision could vary as much as 10% depending on which rater's judgments they used as the gold standard. For some grammatical errors such as subject-verb agreement, where rules are clearly defined, it may be acceptable to use just one rater. But for usage errors, the rules are less clearly defined and two native speakers can have very different judgments of what is acceptable. One way to address this is by aggregating a multitude of judgments for each preposition and treating this as the gold standard, however such a tactic has been impractical due to time and cost limitations.

While annotation is a problem in this field, comparing one system to another has also been a major issue. To date, *none* of the preposition and article error detection systems in the literature have been evaluated on the same corpus. This is mostly due to the fact that learner corpora are difficult to acquire (and then annotate), but also to the fact that they are

[1] (Tetreault and Chodorow, 2008b) report that it would take 80hrs for one of their trained raters to find and mark 1,000 preposition errors.

Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pages 45–48,
Los Angeles, California, June 2010.

usually proprietary and cannot be shared. Examples include the Cambridge Learners Corpus[2] used in (Felice and Pullman, 2009), and TOEFL data, used in (Tetreault and Chodorow, 2008a). This makes it difficult to compare systems since learner corpora can be quite different. For example, the "difficulty" of a corpus can be affected by the L1 of the writers, the number of years they have been learning English, their age, and also where they learn English (in a native-speaking country or a non-native speaking country). In essence, learner corpora are not equal, so a system that performs at 50% precision in one corpus may actually perform at 80% precision on a different one. Such an inability to compare systems makes it difficult for this NLP research area to progress as quickly as it otherwise might.

In this paper we show that the Amazon Mechanical Turk (AMT), a fast and cheap source of untrained raters, can be used to alleviate several of the evaluation and annotation issues described above. Specifically we show:

- In terms of cost and time, AMT is an effective alternative to trained raters on the tasks of preposition selection in well-formed text and preposition error annotation in ESL text.
- With AMT, it is possible to efficiently collect multiple judgments for a target construction. Given this, we propose a new method for evaluation that finally allows two systems to be compared to one another even if they are tested on different corpora.

2 Amazon Mechnical Turk

Amazon provides a service called the Mechanical Turk which allows requesters (companies, researchers, etc.) to post simple tasks (known as Human Intelligence Tasks, or HITs) to the AMT website for untrained raters to perform for payments as low as $0.01 in many cases (Sheng et al., 2008). Recently, AMT has been shown to be an effective tool for annotation and evalatuation in NLP tasks ranging from word similarity detection and emotion detection (Snow et al., 2008) to Machine Translation quality evaluation (Callison-Burch, 2009). In these cases, a handful of untrained AMT workers (or Turkers) were found to be as effective as trained raters, but with the advantage of being considerably faster and less expensive. Given the success of using AMT in other areas of NLP, we test whether we can leverage it for our work in grammatical error detection, which is the focus of the pilot studies in the next two sections.

The presence of a *gold standard* in the above papers is crucial. In fact, the usability of AMT for text annotation has been demostrated in those studies by showing that non-experts' annotation converges to the gold standard developed by expert annotators. However, in our work we concentrate on tasks where there is no single gold standard, either because there are multiple prepositions that are acceptable in a given context or because the conventions of preposition usage simply do not conform to strict rules.

3 Selection Task

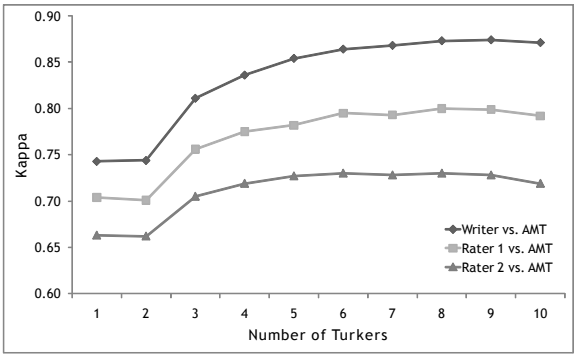

Figure 1: Error Detection Task: Reliability of AMT as a function of number of judgments

Typically, an early step in developing a preposition or article error detection system is to test the system on well-formed text written by native speakers to see how well the system can predict, or select, the writer's preposition given the context around the preposition. (Tetreault and Chodorow, 2008b) showed that trained human raters can achieve very high agreement (78%) on this task. In their work, a rater was shown a sentence with a target preposition replaced with a blank, and the rater was asked to select the preposition that the writer may have used. We replicate this experiment not with trained raters but with the AMT to answer two research questions: 1. Can untrained raters be as effective as trained

[2]http://www.cambridge.org/elt

raters? 2. If so, how many raters does it take to match trained raters?

In the experiment, a Turker was presented with a sentence from Microsoft's Encarta encyclopedia, with one preposition in that sentence replaced with a blank. There were 194 HITs (sentences) in all, and we requested 10 Turker judgments per HIT. Some Turkers did only one HIT, while others completed more than 100, though none did all 194. The Turkers' performance was analyzed by comparing their responses to those of two trained annotators and to the Encarta writer's preposition, which was considered the gold standard in this task. Comparing each trained annotator to the writer yielded a kappa of 0.822 and 0.778, and the two raters had a kappa of 0.742. To determine how many Turker responses would be required to match or exceed these levels of reliability, we randomly selected samples of various sizes from the sets of Turker responses for each sentence. For example, when samples were of size N = 4, four responses were randomly drawn from the set of ten responses that had been collected. The preposition that occurred most frequently in the sample was used as the Turker response for that sentence. In the case of a tie, a preposition was randomly drawn from those tied for most frequent. For each sample size, 100 samples were drawn and the mean values of agreement and kappa were calculated. The reliability results presented in Table 1 show that, with just three Turker responses, kappa with the writer (top line) is comparable to the values obtained from the trained annotators (around 0.8). Most notable is that with ten judgments, the reliability measures are much higher than those of the trained annotators. [3]

4 Error Detection Task

While the previous results look quite encouraging, the task they are based on, preposition selection in well-formed text, is quite different from, and less challenging than, the task that a system must perform in detecting errors in learner writing. To examine the reliability of Turker preposition error judgments, we ran another experiment in which Turkers were presented with a preposition highlighted in a sentence taken from an ESL corpus, and were instructed to judge its usage as either *correct*, *incorrect*, or *the context is too ungrammatical to make a judgment*. The set consisted of 152 prepositions in total, and we requested 20 judgments per preposition. Previous work has shown this task to be a difficult one for trainer raters to attain high reliability. For example, (Tetreault and Chodorow, 2008b) found kappa between two raters averaged 0.630.

Because there is no gold standard for the error detection task, kappa was used to compare Turker responses to those of three trained annotators. Among the trained annotators, inter-kappa agreement ranged from 0.574 to 0.650, for a mean kappa of 0.606. In Figure 2, kappa is shown for the comparisons of Turker responses to each annotator for samples of various sizes ranging from N = 1 to N = 18. At sample size N = 13, the average kappa is 0.608, virtually identical to the mean found among the trained annotators.

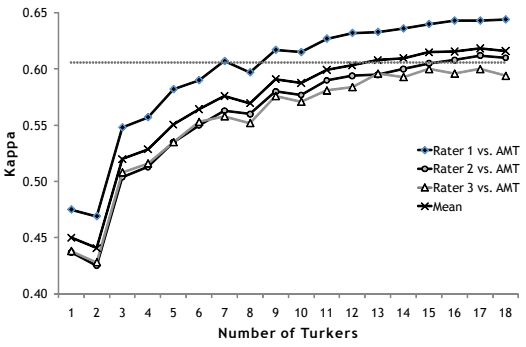

Figure 2: Error Detection Task: Reliability of AMT as a function of number of judgments

5 Rethinking Evaluation

We contend that the Amazon Mechanical Turk can not only be used as an effective alternative annotation source, but can also be used to revamp evaluation since multiple judgments are now easily acquired. Instead of treating the task of error detection as a "black or white" distinction, where a preposition is either correct or incorrect, cases of preposition use can now be grouped into bins based on the level of agreement of the Turkers. For example, if 90% or more judge a preposition to be an error,

[3] We also experimented with 50 judgments per sentence, but agreement and kappa improved only negligibly.

Task	# of HITs	Judgments/HIT	Total Judgments	Cost	Total Cost	# of Turkers	Total Time
Selection	194	10	1,940	$0.02	$48.50	49	0.5 hours
Error Detection	152	20	3,040	$0.02	$76.00	74	6 hours

Table 1: AMT Experiment Statistics

the high agreement is strong evidence that this is a clear case of an error. Conversely, agreement levels around 50% would indicate that the use of a particular preposition is highly contentious, and, most likely, it should not be flagged by an automated error detection system.

The current standard method treats all cases of preposition usage equally, however, some are clearly harder to annotate than others. By breaking an evaluation set into agreement bins, it should be possible to separate the "easy" cases from the "hard" cases and report precision and recall results for the different levels of human agreement represented by different bins. This method not only gives a clearer picture of how a system is faring, but it also ameliorates the problem of cross-system evaluation when two systems are evaluated on different corpora. If each evaluation corpus is annotated by the same number of Turkers and with the same annotation scheme, it will now be possible to compare systems by simply comparing their performance on each respective bin. The assumption here is that prepositions which show X% agreement in corpus A are of equivalent difficulty to those that show X% agreement in corpus B.

6 Discussion

In this paper, we showed that the AMT is an effective tool for annotating grammatical errors. At a fraction of the time and cost, it is possible to acquire high quality judgments from *multiple* untrained raters without sacrificing reliability. A summary of the cost and time of the two experiments described here can be seen in Table 1. In the task of preposition selection, only three Turkers are needed to match the reliability of two trained raters; in the more complicated task of error detection, up to 13 Turkers are needed. However, it should be noted that these numbers can be viewed as upper bounds. The error annotation scheme that was used is a very simple one. We intend to experiment with different guidelines and instructions, and to screen (Callison-Burch, 2009) and weight Turkers' responses (Snow et al., 2008), in order to lower the number of Turkers required for this task. Finally, we will look at other errors, such as articles, to determine how many Turkers are necessary for optimal annotation.

Acknowledgments

We thank Sarah Ohls and Waverely VanWinkle for their annotation work, and Jennifer Foster and the two reviewers for their comments and feedback.

References

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *EMNLP*.

Rachele De Felice and Stephen G. Pullman. 2009. Automatic detection of preposition errors in learner writing. *CALICO Journal*, 26(3).

Michael Gamon, Jianfeng Gao, Chris Brockett, Alex Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for esl error correction. In *Proceedings of IJCNLP*, Hyderabad, India, January.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12:115–129.

Victor Sheng, Foster Provost, and Panagiotis Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceeding of ACM SIGKDD*, Las Vegas, Nevada, USA.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*.

Joel R. Tetreault and Martin Chodorow. 2008a. The ups and downs of preposition error detection in ESL writing. In *COLING*.

Joel Tetreault and Martin Chodorow. 2008b. Native Judgments of non-native usage: Experiments in preposition error detection. In *COLING Workshop on Human Judgments in Computational Linguistics*.