How Far are We from Fully Automatic High Quality Grammatical Error Correction?

Christopher Bryant
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
bryant@comp.nus.edu.sg

Hwee Tou Ng
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nght@comp.nus.edu.sg

Abstract

In this paper, we first explore the role of inter-annotator agreement statistics in grammatical error correction and conclude that they are less informative in fields where there may be more than one correct answer. We next created a dataset of 50 student essays, each corrected by 10 different annotators for all error types, and investigated how both human and GEC system scores vary when different combinations of these annotations are used as the gold standard. Upon learning that even humans are unable to score higher than 75% $F_{0.5}$, we propose a new metric based on the ratio between human and system performance. We also use this method to investigate the extent to which annotators agree on certain error categories, and find that similar results can be obtained from a smaller subset of just 10 essays.

1 Introduction

Interest in grammatical error correction (GEC) systems has grown considerably in the past few years, thanks mainly to the success of the recent Helping Our Own (HOO) (Dale and Kilgarriff, 2011; Dale et al., 2012) and Conference on Natural Language Learning (CoNLL) (Ng et al., 2013; Ng et al., 2014) shared tasks. Despite this increasing attention, however, one of the most significant challenges facing GEC today is the lack of a robust evaluation practice. In fact Chodorow et al. (2012) even go as far to say that it is sometimes "hard to draw meaningful comparisons between different approaches, even when they are evaluated on the same corpus."

One of the reasons for this is that, traditionally, system performance has only ever been evaluated against the gold standard annotations of a single native speaker (rarely, two native speakers). As such, system output is not actually scored on the basis of grammatical acceptability alone, but rather is also constrained by the idiosyncrasies of the particular annotators.

The obvious solution to this problem would be to compare systems against the gold standard annotations of multiple annotators, in an effort to dilute the effect of individual annotator bias, however creating manual annotations is often considered too time consuming and expensive. In spite of this, while other studies have instead elected to use crowdsourcing to produce multiply-corrected annotations, often concerning only a limited number of error types (Madnani et al., 2011; Pavlick et al., 2014; Tetreault et al., 2014), one of the main contributions of this paper is the provision of a dataset of 10 human expert annotations, annotated in the tradition of CoNLL-2014, that is moreover annotated for all error types.[1]

With this new dataset, we have, for the first time, been able to compare system output against the gold standard annotations of a larger group of human annotators, in a realistic grammar checking scenario, and consequently been able to quantify the extent to which additional annotators affect system performance. Additionally, we also noticed that some annotators tend to agree on certain error categories more than others and so attempt to explain this.

In light of the results, we also explore how human annotators themselves compare against the combined annotations of the remaining annotators and thus calculate an upper bound $F_{0.5}$ score for the given dataset and number of annotators; e.g., if one human versus nine other humans is only able to score a maximum of 70% $F_{0.5}$, then it is unreasonable to expect a machine to do better. For this reason, we propose a more informative method of

[1] http://www.comp.nus.edu.sg/~nlp/sw/10gec_annotations.zip

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 697–707, Beijing, China, July 26-31, 2015.

evaluating a system based on the ratio of that system's $F_{0.5}$ score against the equivalent human $F_{0.5}$ score.

Section 2 contains an overview of some of the latest research in both GEC and SMT that makes use of IAA statistics. Section 3 shows an example sentence from our dataset and qualitatively analyses how individual annotator bias affects their choice of corrections. Section 4 describes the data collection process and presents some preliminary results. Section 5 discusses the main quantitative results of the paper, formalizing the formulas used and introducing the more informative method of ratio scoring for GEC, while Section 6 summarizes the results from our additional experiments on category agreement and essay subsets. Section 7 concludes the paper.

2 Inter-Annotator Agreement (IAA)

Whenever we discuss multiple annotators, researchers invariably raise the issue of inter-annotator agreement (IAA), or rather the extent to which annotators agree with each other. This is because data which shows a higher level of agreement is often believed to be in some way more reliable than data which has a lower agreement score. Within GEC, agreement has often been reported in terms of Cohen's-κ (Cohen, 1960), although other agreement statistics could also be used.[2]

In the rest of this section, however, we wish to challenge the use of IAA statistics in GEC and question their value in this field. Specifically, while IAA statistics may be informative in areas where items can be classified into single, well-defined categories, such as in part-of-speech tagging, we argue that they are less well-suited to GEC and SMT, where there is often more than one correct answer. For example, two annotators may correct or translate a given sentence in two completely different yet valid ways, but IAA statistics are only able to interpret the alternative answers as disagreements.

2.1 Inter-Annotator Agreement in GEC

One important study that made use of κ as a measure of agreement between raters is by Tetrault and Chodorow (2008) (also in Tetreault et al. (2014)), who asked two native English speakers to insert a missing preposition into 200 randomly chosen, well-formed sentences from which a single preposition had been removed.

Despite the simplicity of this correction task, the authors reported κ-agreement of just 0.7, noting that in cases where the raters disagreed, their disagreements were often "licensed by context" and thus actually "acceptable alternatives". This led them to conclude that they would "expect even more disagreement when the task is preposition error detection in 'noisy' learner texts" and, by extension, imply that detection of *all* error types in 'noisy' texts would show more disagreement still.

The most important question to ask then, as a result of this study, is whether low κ-scores in 'noisy' texts are truly indicative of real disagreement, or whether, as in this preposition test, the disagreement is actually the result of multiple correct answers, and therefore not disagreement at all.

In a related study, and aware of the fact that there are often multiple ways to correct individual words in sentence, Rozovskaya and Roth (2010) instead chose to compute agreement at the sentence level. Specifically, three raters were asked simply to decide whether they thought 200 sentences were correct or not.

This time, despite operating at the more general sentence level, the authors reported κ scores of just 0.16, 0.4 and 0.23, surmising that "the low numbers reflect the difficulty of the task and the variability of the native speakers' judgments about acceptable usage." If that is the case, then true disagreement may be indistinguishable from native variability, and we should be wary of using IAA statistics as a measure of agreement or evaluation in GEC.

2.2 Inter-Annotator Agreement in SMT

In fact, the issues regarding the reliability of IAA metrics are not unique to GEC and we can also draw a parallel with the field of statistical machine translation (SMT). In the same way that there is often more than one way to correct a sentence in GEC, it is also well known that there is often more than one way to translate a sentence in SMT.

Nevertheless, while several papers have successfully discussed ways to minimize annotator bias effects in SMT (Snover et al., 2006; Madnani et al., 2008), IAA metrics such as κ still unhelpfully play a role in the field and have, for example, been reported almost every year in the Workshop on Machine Translation (WMT) conference.

[2] See Hayes and Krippendorff (2007) or Artstein and Poesio (2008) for the pros and cons of different IAA metrics.

Source:	**To put it in the nutshell**, I believe that people should **have the obligation** to tell their relatives about **the** genetic **testing result** for the good of their health.
A1	To put it in **a** nutshell, I believe that people should **be obliged** to tell their relatives about **their** genetic **test results** for the good of their health.
A2	**In a nutshell**, I believe that people should have **an** obligation to tell their relatives about the genetic testing result for the good of their health.
A3	**In summary**, I believe that people should have the obligation to tell their relatives about the genetic testing result for the good of their health.
A4	**In a nutshell**, I believe that people should **be obligated** to tell their relatives about the genetic testing result for the good of their health.
A5	To put it in **a** nutshell, I believe that people should **be obligated** to tell their relatives about the genetic testing **results** for the good of their health.
A6	To put it in the nutshell, I believe that people should have **an** obligation to tell their relatives about **their** genetic **test results** for the good of their health.
A7	To put it in **a** nutshell, I believe that people should have the obligation to tell their relatives about the genetic testing result for the good of their health.
A8	To put it in **a** nutshell, I believe that people should **be obligated** to tell their relatives about the genetic testing result for the good of their health.
A9	To put it in **a** nutshell, I believe that people should have the obligation to tell their relatives about the genetic **test** result for the good of their health.
A10	To put it in **a** nutshell, I believe that people should have the obligation to tell their relatives about the genetic **test results** for the good of their health.

Table 1: Table showing how each of the 10 annotators edited the same source sentence in Essay 25. The words in the source sentence that were changed are highlighted in bold.

This is in spite of the fact that the average inter-annotator κ score across all language pairs over the past five years has never been higher than 0.4 (Bojar et al., 2014).

One important paper that attempts to explain why IAA metrics score so poorly in SMT is by Lommel et al. (2014), who asked annotators to highlight and categorize sections of automatically translated text they believed to be erroneous. Their results showed that while annotators were often able to agree on the rough locations of errors, they often disagreed as to the specific boundaries of those errors: for instance, given the phrase "had go", some annotators considered just the participle "go" → "gone" to be the minimal error, while others considered the whole verbal unit, "had go" → "had gone", to be the minimal error. Similarly, the authors also noted that annotators sometimes had problems categorizing ambiguous errors which could be classified into more than one error category.

In short, while annotators already vary as to what they consider an error, these observations show that even when they do apparently agree, there is no guarantee that every annotator will define the error in exactly the same terms. This poses a problem for IAA statistics, which rely on an exact match to measure agreement.

Finally, it is also worth mentioning that a related study, by Denkowski and Lavie (2010), suggested that "annotators also have difficulty agreeing with themselves" (shown from *intra*-annotator agreement κ scores of about 0.6), and so we should be especially wary of using IAA metrics to validate datasets that may even be unreliable for a single annotator.

3 Annotator Bias

In an effort to better understand how annotators' judgments might differ, we first carried out a small-scale qualitative analysis on a handful of random sentences corrected by the 10 human annotators in our dataset. One such sentence, and all its various corrections, is shown in Table 1.

It is interesting to note that, for even as short an idiom as "To put it in the nutshell', there are still multiple alternative edits. Although 8 out of the 10 annotators elected to replace the article "the" with "a", among them, A2 and A4 also deleted "To put it" from the expression. Of the remaining 2 annotators, A3 chose to replace the idiom entirely with "In summary", while A6 made no correction at all. Although no correction appears to be unacceptable to the majority of annotators, it is also not completely ungrammatical (just idiomatically awkward) so it may be that A6 has a higher tolerance for this kind of error than the other annotators. Alternatively, there is also always the possibility that, given such a large amount of text to correct, this error was simply overlooked.

Another noteworthy difference is that annotators A1, A4, A5, and A8 all elected to change the

verb "have the obligation" from active to passive, although A1 still disagreed with the others on the form of the participle. Similarly, there is also a great difference of opinion on whether "testing result" should be corrected or not, and if so, how. While half of the annotators left the phrase unchanged, A1, A6, and A10 all changed both words to "test results". Meanwhile, somewhere in between, A5 decided to change "result" to "results", but not "testing" to "test", while, conversely, A9 decided to do the opposite. This would suggest that error correction of even minor phrases falls along a continuum governed by each annotator's natural bias.

Finally, one of the most important results of this qualitative evaluation is that even though all 10 annotators edited the same sentence to a level they deemed grammatical, not one single annotator agreed with another exactly. This fact alone suggests IAA statistics are not a good way to evaluate GEC data and that a more robust agreement metric must take into account the possibility of alternative correct answers.

4 Data Collection

The raw text data in our dataset was originally produced by 25 students at the National University of Singapore (NUS) who were non-native speakers of English. They were asked to write two essays on the topics of genetic testing and social media respectively. All essays were of similar length and quality. This was important because varying the skill level of the essays is likely to further affect the natural bias of the annotators, who may then consistently over- or under-correct essays. These raw essays also formed the basis of the CoNLL-2014 test data (Ng et al., 2014). See Table 2 for some basic statistics on the resulting 50 essays.

The 10 annotators who annotated all 50 essays include: the 2 official annotators of CoNLL-2014, the first author of this paper, and 7 freelancers who were recruited via online recruitment website, Elance.[3] All annotators are native British English speakers, many of whom also have backgrounds in English language teaching, proofreading, and/or Linguistics.

All annotations were made using an online annotation platform, WAMP, especially designed for annotating ESL errors (Dahlmeier et al., 2013). Using this platform, annotators were asked to highlight a minimal error string in the source text, provide an appropriate correction, and then categorize their selection according to the same 28-category error framework used by CoNLL-2014. Before commencing annotation, however, each annotator was given detailed instructions on how to use the tool, along with an explanation of each of the error categories. In cases of uncertainty, annotators were also encouraged to ask questions.

[3] http://www.elance.com

	Total	Average per essay
# Paragraphs	252	5.0
# Sentences	1312	26.2
# Tokens	30144	602.9

Table 2: Statistics for the 50 unannotated essays.

As it was slightly harder to control the quality of the 7 independently recruited annotators via Elance, they were each preliminarily asked to annotate only the first two essays before being given detailed feedback on their work. The main purpose of this feedback was to make sure that they a) understood the error category framework, and b) knew how to deal with more complicated cases such as word insertions, punctuation, etc. Unless it was felt that they had overlooked an obvious error in these first two essays, the feedback did not go so far as to tell annotators what they should and should not highlight in an effort to preserve individual annotator bias.

In all, while the specific time taken to complete annotation of all 50 essays was not calculated, all annotators completed the task over a period of about 3 weeks, at a rate of about 45 minutes per essay.

4.1 Early Observations

To investigate the extent to which different annotators have different biases, we first counted the total number of edits made by each annotator and sorted them by error category (Table 3).

As can be seen, there is quite a difference between the annotator who made the most edits (A1) and the annotator who made the fewest edits (A7), with A1 making more than twice the number of edits as A7. This just goes to show how varied judgments on grammaticality can be. Incidentally, annotators A3 and A7, who are among those who made the fewest edits, were also the two official gold standard annotators in CoNLL-2014.

There is also a large difference between edits in

Category	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Total
ArtOrDet	879	639	443	503	665	620	331	358	390	624	5452
Cit	0	0	0	0	0	1	0	2	0	0	3
Mec	227	376	493	325	411	336	228	733	598	780	4507
Nn	404	290	228	264	360	300	215	254	277	365	2957
Npos	21	21	15	21	31	28	19	25	29	23	233
Others	42	186	49	116	95	43	44	34	125	105	839
Pform	431	52	18	57	30	83	47	53	19	18	808
Pref	4	79	153	18	223	53	96	92	250	180	1148
Prep	755	488	390	421	502	556	211	276	362	459	4420
Rloc–	488	308	199	331	187	244	94	174	296	240	2561
Sfrag	1	5	1	3	1	5	13	2	12	2	45
Smod	1	4	5	0	1	0	0	3	1	1	16
Spar	0	18	24	0	2	11	3	2	8	0	68
Srun	157	38	21	16	17	18	7	15	17	37	343
Ssub	74	54	10	4	25	81	68	21	18	82	437
SVA	162	123	154	95	140	114	105	132	144	144	1313
Trans	248	100	78	147	118	81	93	199	87	95	1246
Um	5	12	42	25	25	12	12	19	7	8	167
V0	137	35	37	50	81	69	31	58	51	85	634
Vform	388	168	91	100	156	125	132	78	122	124	1484
Vm	71	48	37	67	119	24	49	39	4	62	520
Vt	100	209	150	200	82	237	133	234	117	188	1650
Wa	0	1	1	3	1	1	0	2	4	2	15
Wci	623	476	479	446	456	595	340	250	212	346	4223
Wform	126	107	103	150	136	145	77	103	107	81	1135
WOadv	23	48	27	23	61	76	12	94	41	62	467
WOinc	187	67	54	78	53	74	22	24	87	103	749
Wtone	6	30	15	65	38	27	9	10	12	15	227
Total	5560	3982	3317	3528	4016	3959	2391	3286	3397	4231	37667

Table 3: Table showing how many annotations each annotator made in terms of error category. See Ng et al. (2014) Table 1 for a more detailed description of error categories.

terms of category use, with almost half of all edits falling into the categories for article or determiner (ArtOrDet), spelling or punctuation (Mec), preposition (Prep), or word choice (Wci) errors.

5 Quantitative Analysis

In the main phase of experimentation, we first investigated how different numbers of annotators affected the performance of various systems in the context of the CoNLL-2014 shared task. To do this, we downloaded the official system output of all the participating teams[4] and then the *MaxMatch* (M2) Scorer[5] (Dahlmeier and Ng, 2012), which was the official scorer of the previous CoNLL-2013 and CoNLL-2014 shared tasks.

This scorer evaluates a system at the sentence level in terms of correct edits, proposed edits, and gold edits, and uses these to calculate an F-score for each team. When more than one set of gold standard annotations is available, the scorer will calculate F-scores for each alternative gold-standard *sentence* and choose the one from whichever annotator scored the highest. As in CoNLL-2014, we calculate $F_{0.5}$, which weights precision twice as much as recall, because it is more important for a system to be accurate than to correct every possible error. See (Ng et al., 2014) for more details on how $F_{0.5}$ is calculated.

5.1 Pairwise Evaluation

In order to quantify how much the F-score can vary in a realistic grammar checking scenario when there is only one gold standard annotator, we first computed the scores for a participating system vs each annotator in a pairwise fashion. Table 4 hence shows how the top team in CoNLL-2014, CAMB (Felice et al., 2014), performed against each of the 10 human annotators individually.

While Tetrault and Chodorow (2008) and Tetreault et al. (2014) reported a difference of 10% precision and 5% recall between their two individual annotators in their simplified preposition correction task, Table 4 shows this difference can actually be as much as almost 15% precision (A1 vs A7) and 6% recall (A1 vs A3) in a more realistic full scale correction task. This equates to a differ-

[4] http://www.comp.nus.edu.sg/~nlp/conll14st/official_submissions.tar.gz

[5] http://www.comp.nus.edu.sg/~nlp/sw/m2scorer.tar.gz

CAMB	P	R	$F_{0.5}$
A1	39.64	14.06	29.06
A2	35.73	17.35	29.48
A3	35.22	20.29	30.70
A4	32.69	17.88	28.04
A5	35.74	17.26	29.43
A6	35.76	17.73	29.72
A7	24.96	19.62	23.67
A8	29.17	16.92	25.48
A9	32.03	18.28	27.84
A10	35.52	16.26	28.72

Table 4: Table showing the $F_{0.5}$ scores for the top team in CoNLL-2014, CAMB, against each of the 10 annotators individually.

ence of over 7% $F_{0.5}$ (A3 vs A7) and once again shows how varied annotator's judgments can be.

5.2 All Combinations

5.2.1 Human vs Human

Whereas previously we could only calculate $F_{0.5}$ scores on a system vs human basis, when there are two or more annotators, we can also calculate scores on a human vs human basis. In fact, as the number of annotators increases, we can also start to calculate scores against different combinations of gold standard annotations.[6]

To give an example, since we have 10 annotators, a subset of these annotators, say annotators a2–a8, could be chosen as the gold standard annotations. We could then evaluate how each of the remaining annotators (i.e., annotator a1, a9, and a10) performs against this gold standard, by computing the M2 score for annotator a1 against annotators a2–a8, annotator a9 against annotators a2–a8, and annotator a10 against annotators a2–a8. We then average these 3 M2 scores, to determine how, on average, an annotator performs when measured against gold standard annotators a2–a8.

It is worth reiterating, however, that when more than one annotator is used as the gold standard, the M2 scorer will choose whichever annotator for the given *sentence* produces the highest F-score; i.e., if a2–a8 are the gold standard and we want to compute the F-score for a9, the M2 scorer will compute a9 vs a2, a9 vs a3, . . . , a9 vs a8 separately *for each sentence*, and choose the highest.

[6]Note that by combinations of annotators, we mean simply that the M2 scorer has access to a larger number of alternative gold standard corrections; we do not attempt to merge annotations in any way.

The above calculations can be formalized as Equation 1:

$$g(X) = \frac{1}{|A| - |X|} \sum_{a \in A \setminus X} f(a, X) \qquad (1)$$

where A is the set of all annotators ($|A| = 10$ in our case) and X is a non-empty and proper subset of A, denoting the set of annotators chosen to be in the gold standard. The function $f(a, X)$ is the score computed by the M2 scorer to evaluate annotator a against each set of gold standard annotators X. $g(X)$ is thus the average M2 scores for the remaining annotators against the input gold standard combination X.

So far, in our example, we have chosen annotators a2–a8 to be the gold standard. There are, however, many other different ways of choosing 7 annotators to serve as the gold standard. For example, we could have chosen { a1, a2, . . . , a7 }, { a1, a3, a4, . . . , a8 }, etc. In fact, there are $\binom{10}{7} = 120$ different combinations of 7 annotators. As such, we can also compute how an individual human annotator performs when measured against any combination of 7 gold standard annotators, by averaging these 120 M2 scores. The above calculation is formalized in the general case in Equation 2:

$$h_i = \frac{1}{\binom{|A|}{|X|}} \sum_{X:|X|=i} g(X) \qquad (2)$$

where $\binom{|A|}{|X|}$ is the binomial coefficient for $|A|$ choose $|X|$ and $1 \leq i < |A|$. The function $g(X)$ is defined in Equation 1.

The resulting h_i values are hence the average $F_{0.5}$ scores achieved by any human against any combination of i other humans, and so, in some ways, also represent the upper bound of human performance on the current dataset. The specific values for h_i are shown in the second column of Table 5.

5.2.2 Caveat

One caveat regarding this method is that the number of *all* possible combinations of annotators is of the order $2^{|A|}$, which quickly becomes computationally expensive for large values of $|A|$. Fortunately however, in a realistic GEC evaluation scenario, it is only the last row of Table 5 that we are most interested in, and so it is actually only necessary to calculate a much more manageable $\binom{|A|}{|A|-1}$ gold standard combinations, which is conveniently

Gold Annotators (*i*)	Human (h_i)	AMU		CAMB		CUUI	
	Avg $F_{0.5}$	Avg $F_{0.5}$	Ratio	Avg $F_{0.5}$	Ratio	Avg $F_{0.5}$	Ratio
1	45.91	24.20	52.71%	28.22	61.46%	26.76	58.29%
2	56.68	33.47	59.05%	37.77	66.64%	36.04	63.59%
3	61.83	38.35	62.03%	42.68	69.03%	40.76	65.92%
4	65.05	41.53	63.85%	45.87	70.51%	43.77	67.29%
5	67.33	43.84	65.11%	48.17	71.54%	45.94	68.23%
6	69.07	45.62	66.06%	49.93	72.29%	47.60	68.92%
7	70.45	47.06	66.80%	51.34	72.87%	48.94	69.46%
8	71.60	48.26	67.40%	52.50	73.32%	50.05	69.89%
9	72.58	49.28	67.90%	53.47	73.67%	50.99	70.25%

Table 5: Table showing average human $F_{0.5}$ scores over all combinations of $1 \leq i < 10$ gold annotators compared to the same averages for the top 3 systems in CoNLL-2014, and the ratio percentage of each team's average score versus the human average score.

equal to the total number of annotators. We only compute all combinations here in order to quantify, for the first time, how much each additional annotator affects performance.

5.2.3 System vs Human

In addition to calculating scores on a human vs human basis, we also calculated the F-scores for the top three CoNLL-2014 teams, AMU (Junczys-Dowmunt and Grundkiewicz, 2014), CAMB (Felice et al., 2014), and CUUI (Rozovskaya et al., 2014), versus all the combinations of humans (Equation 3).

$$s_i = \frac{1}{\binom{|A|}{|X|}} \sum_{X:|X|=i} f(s, X) \quad (3)$$

Specifically, $s \in S$, where S is the set of all three shared task systems, i.e., {AMU, CAMB, CUUI}, and $f(s, X)$ is the same function in Equation 1 which is the score computed by the M2 scorer to evaluate system s against the set of annotators X chosen to be in the gold standard. The average $F_{0.5}$ scores for each of the team's systems versus increasing numbers of i annotators are also shown in Table 5.

We notice from these scores that, as expected, both system and human performance increases as more annotators are used in a gold standard. We do now, however, have data that quantifies exactly how much each additional annotator affects the score. This effect can be more clearly seen in Figure 1.

It is important to note, however, that even with 9 annotators, human output itself does not reach close to 100% $F_{0.5}$ and instead, the difference between the systems and the humans is about 20% $F_{0.5}$. Furthermore, the curves for humans and systems also remain roughly parallel, suggesting human corrections gain as much benefit as system corrections from larger sets of gold standard annotations.

5.3 Ratio Scoring

In light of the above observation that even humans vs humans are unable to score 100% $F_{0.5}$, it thus seems unreasonable to expect machines to do the same. As such, we propose that it is much more informative to score system output against the average performance of humans instead of against the theoretical maximum score. The ratio values for the three CoNLL-2014 teams against the human gold standards of various sizes are hence also reported in Table 5. The most important thing to note is that these figures are not only much higher than the low $F_{0.5}$ values currently reported in the literature, they are also more representative of the state of the art. For instance, it is highly significant that we can report that the top system in CoNLL-2014, CAMB, is actually able to perform 73% as reliably as a human, which suggests GEC may actually be a more viable technology than was previously thought.

6 Additional Experiments

6.1 Error Categories

As well as carrying out experiments at the system level, we also carried out similar experiments at the error category level. More specifically, we recalculated the values of Equation 1 and 2 for cases where the set of annotations consisted of only a

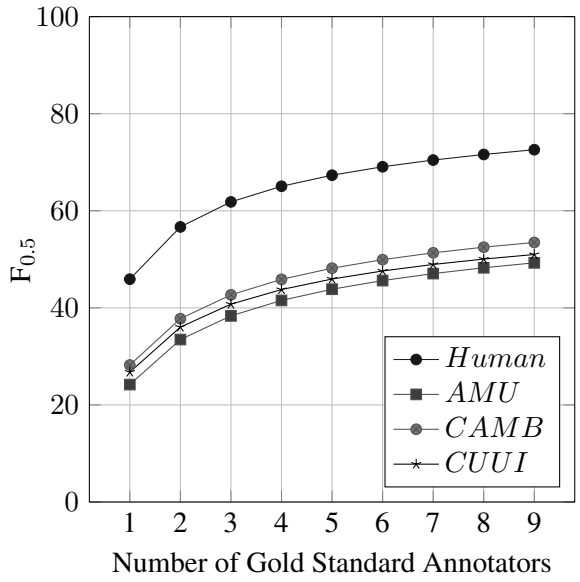

Figure 1: Graph showing how average $F_{0.5}$ scores for humans and systems increase as the number of gold standard annotators also increases (*all error types, 50 Essays*).

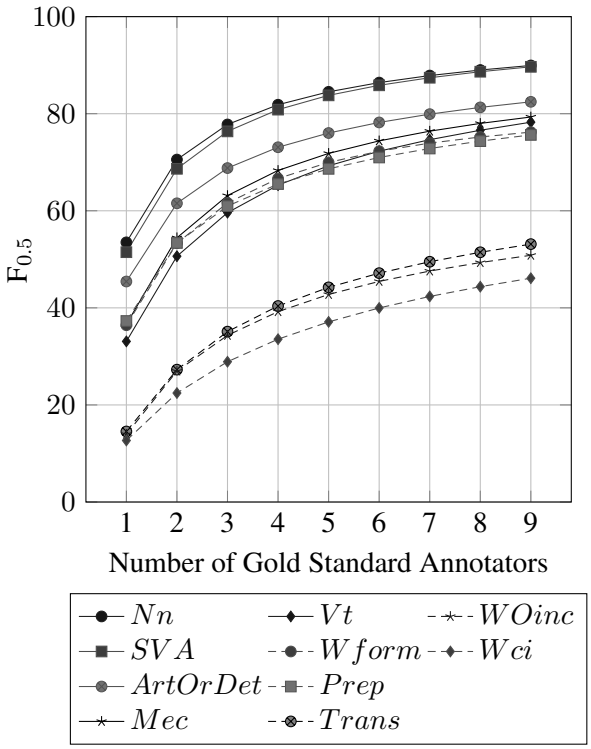

Figure 2: Graph showing how average $F_{0.5}$ scores for various error categories increase as the number of gold standard annotators also increases (*50 essays*). Calculations based on human annotations only.

single specific error type. Since the participating teams in CoNLL-2014 were not asked to classify the type of errors their systems corrected, we were only able to calculate these new values using the 10 sets of human annotations.

Like Figure 1, we can see from Figure 2 that the $F_{0.5}$ performance of individual error types increases diminishingly as the number of annotators in the gold standard also increases. More importantly, however, we notice that some error types achieve much higher scores than others, which suggests some annotators agree on certain categories more than others.

In particular, noun number (Nn) and subject-verb agreement (SVA) errors achieve the highest scores, at just under 90% $F_{0.5}$, which is also not far from the 100% $F_{0.5}$ that would be achieved if we had gold standard answers for all possible alternative corrections of this type. The most likely reason for this is that, as the correction of these error types typically only involves the addition or removal of an -s suffix, i.e., a minor change in number morphology, there is very little room for annotators to disagree.

In contrast, the next highest category, article and determiner errors (ArtOrDet), has a slightly larger confusion set, {the, a/an, ϵ}, which may account for the slightly lower score. Similarly, the next group of error categories, spelling and punctuation (Mec), verb tense (Vt), and word form (Wform), which all often involve a similar type of edit operation to a word lemma, likewise have slightly larger confusion sets that include a larger variety of possible morphological inflections. It is likely that the next category, prepositions (Prep), also has a confusion set of a similar size.

The last three categories, conjunctions (all-types) (Trans), word order (WOinc) and word choice (Wci), are all notable because they perform significantly worse than the hitherto mentioned categories. The main reason for this is that these error types all typically have a scope much larger than most other categories in that they often involve changes at the structural or semantic level; e.g., changing an active to a passive or choosing a synonym. For this reason, there are often many more alternative ways to correct them, meaning they are also much more likely to be affected by annotator bias.

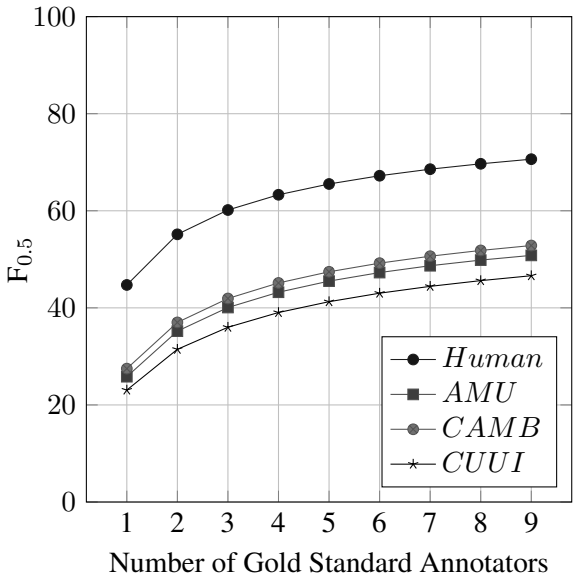

Figure 3: Graph showing how average $F_{0.5}$ scores for humans and systems increase as the number of gold standard annotators also increases (*all error types, 10 Essays*).

6.2 Essay Subsets

Now that we had empirical evidence showing how $F_{0.5}$ scores varied with the number of annotators, an additional question to ask was whether the same trends for 50 essays were also present in a smaller subset of essays. We therefore repeated the main experiment with all error types, but this time used just 10 essays (specifically, essays 1–10) in both the hypothesis and gold standard. The results are shown in Figure 3.

Compared to Figure 1, the most significant difference between these two graphs is that the ranking for AMU and CUUI has changed, although not by much in terms of $F_{0.5}$. The most likely reason for this is that the distribution of error types in the smaller subset of essays is better suited to AMU's more general SMT approach than to CUUI's more targeted classifier based approach. For instance, see Table 9 in Ng et al. (2014) to compare each team's performance on different error types in the CoNLL-2014 shared task.

In other words, while the overall relationship between the system and human scores on 10 and 50 essays remains more or less the same, researchers must be aware that smaller datasets may have more skewed error distributions, which in turn may affect system performance, dependent upon correction strategy. With a balanced test set though, it would seem feasible to carry out future evaluation research on as few as 10 essays (about 6000 words).

7 Conclusion

To summarize, we first showed that 10 individual annotators can all correct the same sentence in 10 different ways, yet also all produce valid alternatives. This implies that inter-annotator agreement statistics, which rely on exact matching, are not well-suited to grammatical error correction, because it may not be the case that annotators truly disagree, but rather that they have a bias towards a particular type of alternative answer.

We next showed that, as has long been suspected, increasing the number of annotators in the gold standard also leads to an increase in $F_{0.5}$, although at a diminishing rate. This data can be used to help researchers decide how many gold standard annotations should be used in GEC evaluation.

The main result of this paper however, is that by computing scores for human against human, we determined that it is *not* true that any human correction is able to score 100% $F_{0.5}$. Instead, we found that the human upper bound is roughly 73% $F_{0.5}$ and that the top 3 teams from CoNLL-2014 actually perform, on average, between 67-73% as reliably as this human upper bound. This result is highly significant, because it suggests GEC systems may actually be more viable than their previously low $F_{0.5}$ scores would suggest.

In addition to the above, we also found that humans tend to agree on some error categories more than others, and suggest that one of the main reasons for this concerns the size of the confusion set of the particular error type.

Finally, not only are we making the corrections by 10 annotators of all 50 essays available with this paper, we also showed that the trends found in the data are also consistent with the annotations of just 10 essays, allowing future research to be conducted on much less text.

Acknowledgments

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 2 grant MOE2013-T2-1-150. We would also like to thank the three anonymous reviewers for their comments.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel R. Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *COLING*, pages 611–628.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *HLT-NAACL*, pages 568–572.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, USA.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. Helping Our Own: HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–62.

Michael Denkowski and Alon Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. *Proceedings of AMTA*.

Mariano Felice, Zheng Yuan, Øistein E Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33.

Arle Richard Lommel, Maja Popovic, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? Investigating the value of paraphrased reference translations in parameter optimization. *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, October*.

Nitin Madnani, Martin Chodorow, Joel R. Tetreault, and Alla Rozovskaya. 2011. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 508–513.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel R. Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. ACL.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, USA. ACL.

Ellie Pavlick, Rui Yan, and Chris Callison-Burch. 2014. Crowdsourcing for grammatical error correction. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW Companion '14, pages 209–212, New York, NY, USA. ACM.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. 2014. The Illinois-Columbia system in the CoNLL-2014 shared task. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 34–42.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the Association for Machine Transaltion in the Americas*.

Joel R. Tetrault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *COLING Workshop on Human Judgments in Computational Linguistics*, pages 24–32, Manchester, UK.

Joel R. Tetreault, Martin Chodorow, and Nitin Madnani. 2014. Bucking the trend: improved evaluation and annotation practices for ESL error detection systems. *Language Resources and Evaluation*, 48(1):5–31.