Human Evaluation of Grammatical Error Correction Systems

Roman Grundkiewicz[1] and **Marcin Junczys-Dowmunt**[1] and **Edward Gillian**[2]

[1]Information Systems Laboratory
Faculty of Mathematics and Computer Science, Adam Mickiewicz University
{romang,junczys}@amu.edu.pl

[2]Faculty of English, Adam Mickiewicz University
egillian@pwsz.pl

Abstract

The paper presents the results of the first large-scale human evaluation of automatic grammatical error correction (GEC) systems. Twelve participating systems and the unchanged input of the CoNLL-2014 shared task have been reassessed in a WMT-inspired human evaluation procedure. Methods introduced for the Workshop of Machine Translation evaluation campaigns have been adapted to GEC and extended where necessary. The produced rankings are used to evaluate standard metrics for grammatical error correction in terms of correlation with human judgment.

1 Introduction

The field of automatic grammatical error correction (GEC) has seen a number of shared tasks of different scope and for different languages. The most impactful were the CoNLL-2013 and CoNLL-2014 (Ng et al., 2013; Ng et al., 2014) shared tasks on Grammatical Error Correction for ESL (English as a second language) learners. They were preceded by the HOO shared tasks (Dale and Kilgarriff, 2011; Dale et al., 2012). Shared tasks for other languages took place as well, including the QALB workshops for Arabic (Mohit et al., 2014) and NLP-TEA competitions for Chinese. These tasks use automatic metrics to determine the quality of the participating systems.

However, these efforts pale in comparison to competitions organized in other fields, e.g. during the annual Workshops for Machine Translation (WMT). It is a central idea of the WMTs that automatic measures of machine translation quality are an imperfect substitute for human assessments. Therefore, manual evaluation of the system outputs are conducted and their results are reported as the final rankings of the workshops. These human evaluation campaigns are an important driving factor for the advancement of MT and produce insightful "by-products", such as a huge number of human assessments of machine translation outputs that have been used to evaluate automatic metrics.

We believe that the unavailability of this kind of quality assessment may stall the development of GEC, as all the shared tasks and the entire field have to cope with an inherent uncertainty of their methods and metrics. We hope to make a step towards alleviating this lack of confidence by presenting the results of the first[1] large-scale human evaluation of automatic grammatical error correction systems submitted to the CoNLL-2014 shared task. Most of our inspiration is drawn from the recent WMT edition (Bojar et al., 2014) and its metrics task (Macháček and Bojar, 2014).

We also provide an analysis of correlation between the standard metrics in GEC and human judgment and show that the commonly used parameters for standard metrics in the shared task may not be optimal. The uncertainty about metrics quality leads to proposals of new metrics, with Felice and Briscoe (2015) being a recent example. Based on human judgments we can show that this proposed metric maybe less useful than hoped.

2 Evaluation of GEC systems

Madnani et al. (2011) addresses two problems of GEC evaluation: 1) a lack of informative metrics and 2) an inability to directly compare the performance of systems developed by different researchers. Two evaluation methodologies are presented, both based on crowdsourcing which are used to grade types of errors rather than system performance as presented in this work. Chodorow et al. (2012) draw attention to the many evalua-

[1] During the camera-ready preparation phase, we learned about similar research by Napoles et al. (2015). After contacting the authors, it was agreed to treat both works as fully concurrent. Future work will compare the results.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 461–470,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

tion issues in error detection which make it hard to compare different approaches. The lack of consensus is due to the nature of the error detection task. The authors argue that the choice of the metric should take into account factors such as the skew of the data and the application that the system is used for.

The most recent addition is Felice and Briscoe (2015) who present a novel evaluation method for grammatical error correction that scores systems in terms of improvement on the original text.

3 The CoNLL-2014 shared task

The goal of the CoNLL-2014 shared task (Ng et al., 2014) was to evaluate algorithms and systems for automatically correcting grammatical errors in English essays written by second language learners of English. Training and test data was annotated with 28 error types. Participating teams were given training data with manually annotated corrections of grammatical errors and were allowed to use publicly available resources for training.

Twenty-five student non-native speakers of English were recruited to write essays to be used as test data. Each student wrote two essays. The 50 test essays were error-annotated by two English native speakers. The essays and error annotations were made available after the task. The MaxMatch (M^2) scorer (Dahlmeier and Ng, 2012) has been used as the official shared task evaluation metric.

4 Data collection

4.1 Sampling sentences for evaluation

The system outputs of the CoNLL-2014 shared task serve as evaluation data. The test set consists of 1312 sentences, there are twelve system outputs available. The thirteenth participant NARA is missing from this set. However, in GEC evaluation there is also the input to consider. Often system outputs are equal to the unmodified input, as it is most desirable if there are in fact no errors. We include INPUT as the thirteenth system.

Due to the small number of modifications that GEC systems apply to the input, there is not only a large overlap with the input, but also among all systems (Figure 1). If we sample systems uniformly, we lose easily obtainable pairwise judgments for systems with the same output, and if we collapse before sampling we introduce a strong bias towards ties. To counter that bias, we abandon uniform sampling of test set sentences and use instead a parametrized distribution that favors diverse sets of outputs.

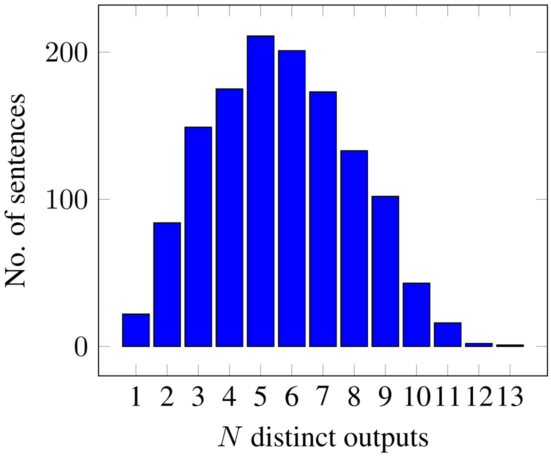


Figure 1: Frequencies of distinct corrected sentences produced by 13 systems per input sentence.

The probability p_i for a set of outputs O_i is calculated as follows: N is the number of systems to be evaluated, M is the maximum number of sentences presented to the evaluator in a single ranking (we use $M = 5$). The set of system outputs to be evaluated $E = \{O_1, \ldots, O_n\} \quad \forall_{1 \le i \le n} |O_i| = N$, consists of n $(= 1312)$ sets O_i of N output sentences each. Every sentence in O_i can overlap with other sentences multiple times, so for each set O_i we define the corresponding multiset of multiplicities U_i, such that $\sum_{u \in U_i} u = N$.

We define $c_i(j)$ as the number of possible ways to choose at most M different sentences that cover j systems for the i-th set of outputs:

$$c_i(j) = \left| \left\{ S \subseteq U_i : |S| \le M \wedge \sum_{u \in S} u = j \right\} \right| .$$

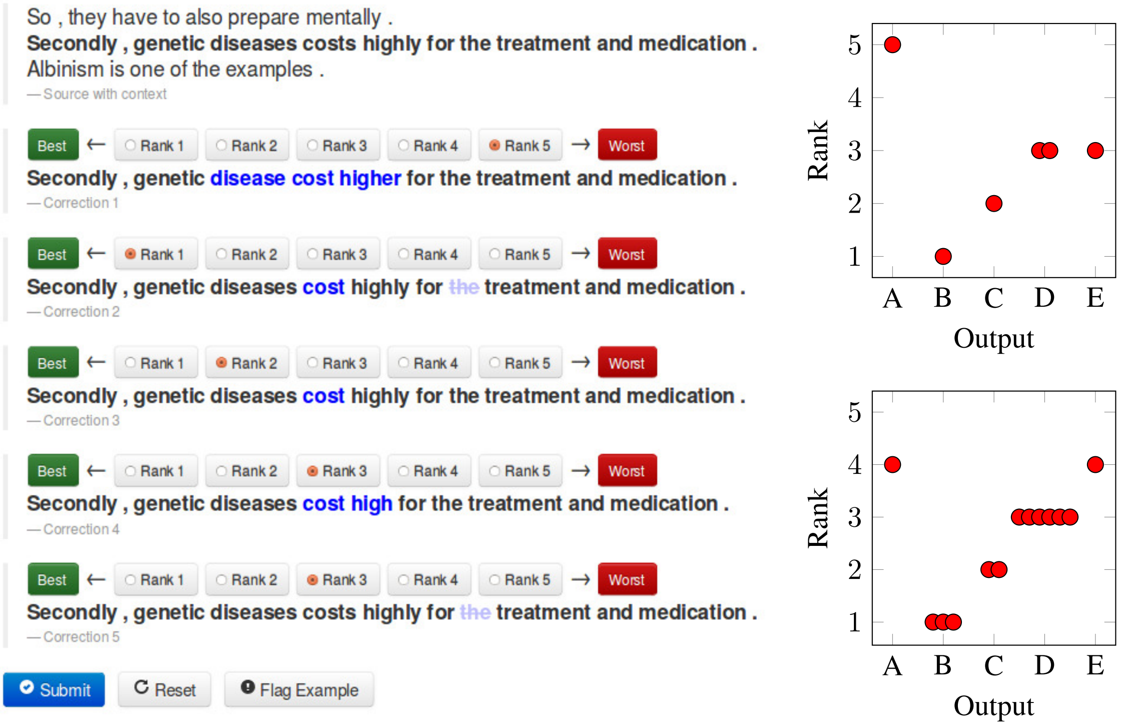
Then the expected number C_i of systems covered by choosing at most M sentences is

$$C_i = \frac{\sum_{j=M}^{N} c_i(j) \cdot j}{\sum_{j=M}^{N} c_i(j)} .$$

The pseudo-probability p'_i of sampling the i-th sentence is defined as

$$p'_i = \frac{\binom{M}{2}}{\binom{C_i}{2}} \quad \text{where} \quad \binom{C_i}{2} = \frac{C_i(C_i - 1)}{2}$$

which is the ratio of pairwise comparisons of M versus C_i different systems. By normalizing over



So , they have to also prepare mentally .
Secondly , genetic diseases costs highly for the treatment and medication .
Albinism is one of the examples .
— Source with context

Secondly , genetic disease cost higher for the treatment and medication .
— Correction 1

Secondly , genetic diseases cost highly for ~~the~~ treatment and medication .
— Correction 2

Secondly , genetic diseases cost highly for the treatment and medication .
— Correction 3

Secondly , genetic diseases cost high for the treatment and medication .
— Correction 4

Secondly , genetic diseases costs highly for ~~the~~ treatment and medication .
— Correction 5

Best ← ○ Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ● Rank 5 → Worst
Best ← ● Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 → Worst
Best ← ○ Rank 1 ● Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 → Worst
Best ← ○ Rank 1 ○ Rank 2 ● Rank 3 ○ Rank 4 ○ Rank 5 → Worst
Best ← ○ Rank 1 ○ Rank 2 ● Rank 3 ○ Rank 4 ○ Rank 5 → Worst
Submit
Reset
Flag Example
Rank
5
4
3
2
1
A
B
C
D
E
Output
Rank
5
4
3
2
1
A
B
C
D
E
Output

(a) Screenshot of Appraise modified for GEC judgment.

(b) Overlapping rankings.

Figure 2: Displayed ranking and corresponding overlapping rankings.

the entire set of output sets we obtain the probability p_i of sampling the i-th set of outputs as

$$p_i = \frac{p'_i}{\sum_{j=1}^{|E|} p'_j}.$$

4.2 Collecting system rankings

The sets of outputs sampled with the described method have been prepared for Appraise (Federmann, 2010) and presented to the judges. Judges were asked to rank sentences from best to worse. Ties are allowed. Judges were aware that the absolute ranks bear no relevance as ranks are later turned into relative pairwise judgments. No notion of "better" or "worse" was imposed by the authors, we relied on the judges to develop their own intuition. All eight judges are English native speakers and have extensive backgrounds in linguistics.

Figure 2a displays a screen shot of Appraise with a judged sentence. Several modifications to the Appraise framework[2] were implemented to account for the specific nature of GEC:

Only the input sentence is displayed (top, bold), no reference correction is given. The input sentence is surrounded by one preceding and one following sentence. Identical corrections are collapsed into one output, system names with the same output are recorded internally. Edited fragments are highlighted, blue for insertions and substitutions, pale blue and crossed-out for deletions.

4.3 Pairwise judgments

As conducted during the WMT campaigns, we turn rankings into sets of relative judgments of the form A>B, A=B, A<B where the lower ranked system scores a win. Absolute ranks and differences are lost. As mentioned above, due to the collapsing of identical outputs we obtain significantly more data than the usual 10 pairs from one ranking with five sentences. Figure 2a contains a ranking with overlapping outputs as displayed in the top graph of Figure 2b. Pairs from within overlaps result in ties, pairs between overlaps are expanded as products, $\binom{6}{2} = 15$ pairwise judgments can be extracted. Greater overlap leads to more pairwise judgments (bottom, $\binom{13}{2} = 78$).

Table 1 lists the full statistics for collected rankings by individual annotators. Unexpanded pairs are WMT-style pairwise judgments before an output A gets split into overlapping systems A_1, A_2, A_3, etc. The large number of ties for expanded pairs is to be expected due to the high overlap

[2] A fork of the original source code with can be found at `https://github.com/snukky/Appraise`

Judge	Ranks	Unexpanded	Expanded
1	400	3525 (1022)	18400 (10166)
2	299	2684 (1099)	13657 (8429)
3	400	3523 (914)	18912 (9684)
4	201	1750 (550)	9478 (5539)
5	349	3099 (766)	17107 (8972)
6	400	3474 (517)	19313 (9209)
7	70	646 (145)	3383 (1593)
8	200	1815 (681)	8848 (5525)
Total	2319	20516 (5694)	109098 (59117)

Table 1: Statistics for collected rankings (Ranks), unexpanded and expanded pairwise judgments, numbers for ties are given in parentheses.

between systems (on average there are only 5.7 unique outputs among 13 systems).

4.4 Inter- and intra-annotator agreement

Again inspired by the WMT evaluation campaigns, we compute annotator agreement as a measure of reliability of the pairwise judgments with Cohen's kappa coefficient (Cohen, 1960):

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}.$$

where $P(A)$ is the proportion of times that annotators agree, and $P(E)$ is the proportion of times that they would agree by chance. κ assumes values from 0 (no agreement) to 1 (perfect agreement).

All probabilities are computed as ratios of empirically counted pairwise judgments. As the judges worked on collapsed outputs, we calculate agreement scores for unexpanded pairs; otherwise, the high overlap would unfairly increase agreement.

$P(A)$ is calculated by examining all pairs of outputs which have been judged by two or more judges, and counting the proportion of times that they agreed that `A<B`, `A=B`, or `A>B`.

$P(E) = P(\texttt{A<B})^2 + P(\texttt{A=B})^2 + P(\texttt{A>B})^2$ is the probability that two judges agree randomly. Intra-annotator agreement as a measure of consistency is calculated for output sets that have been judged more than one time by the same annotator.

The agreement numbers in Table 2 are in the lower range of values reported during WMT. However, it should be noted that judges never saw the repeated outputs within one ranking which probably decreases agreement compared to the MT-specific task.

Agreement	Value	Degree
Inter-annotator	0.29	Weak
Intra-annotator	0.46	Moderate

(a) Inter-annotator and intra-annotator agreement for all judges

	1	2	3	4	5	6	7	8
1	.42	.26	.30	.37	.34	.26	.31	.24
2	–	.30	.25	.28	.23	.20	.10	.20
3	–	–	.50	.35	.44	.34	.46	.26
4	–	–	–	.34	.34	.30	.20	.26
5	–	–	–	–	.60	.36	.34	.32
6	–	–	–	–	–	.44	.35	.25
7	–	–	–	–	–	–	*	*
8	–	–	–	–	–	–	–	.48

(b) Pairwise inter-annotator and intra-annotator agreement per judge. Stars indicate too few overlapping judgements.

Table 2: Inter-annotator and intra-annotator agreement (Cohen's κ) on unexpanded pairwise judgments.

5 Computing ranks

In this section, it is our aim to produce a system ranking from best to worse by computing the average number of times each system was judged better than other systems based on the collected pairwise rankings. While previously introduced methods for producing rankings, total orderings, as well as partial orderings at chosen confidence-levels, can be directly applied to our data, determining which ranking is more accurate turns out to be methodologically and computationally more involved due to the specific nature of GEC outputs.

5.1 Ranking methods

We adapt two ranking methods applied during WMT13 and WMT14 to GEC evaluation: the Expected Wins method and a version of TrueSkill.

Expected Wins. Expected Wins (EW) has been introduced for WMT13 (Bojar et al., 2013) and is based on an underlying model of "relative ability" proposed in Koehn (2012). One advantage of this method is its intuitiveness; the scores reflect the probability that a system S_i will be ranked better than another system that has been randomly chosen from a pool of opponents $\{S_j : j \neq i\}$. Defining the function $\mathrm{win}(A, B)$ as the number of times system A is ranked better than system B, Bojar et

#	System	P	R	$\mathrm{M}^2_{0.5}$
1	CAMB	0.397	0.301	0.373
2	CUUI	0.417	0.248	0.367
3	AMU	0.416	0.214	0.350
4	POST	0.345	0.217	0.308
5	NTHU	0.350	0.188	0.299
6	RAC	0.331	0.149	0.266
7	UMC	0.312	0.144	0.253
8	PKU	0.322	0.136	0.253
9	SJTU	0.301	0.051	0.151
10	UFC	0.700	0.017	0.078
11	IPN	0.112	0.028	0.071
12	IITB	0.307	0.013	0.059
13	INPUT	0.000	0.000	0.000

(a) Official CoNLL-2014 ranking without unpublished NARA system.

#	Score	Range	System
1	0.628	1	AMU
2	0.566	2-3	RAC
	0.561	2-4	CAMB
	0.550	3-5	CUUI
	0.539	4-5	POST
3	0.513	6-8	UFC
	0.506	6-8	PKU
	0.495	7-9	UMC
	0.485	7-10	IITB
	0.463	10-11	SJTU
	0.456	9-12	INPUT
	0.437	11-12	NTHU
4	0.300	13	IPN

(b) Human ExpectedWins ranking (final manual ranking).

#	Score	Range	System
1	0.273	1	AMU
2	0.182	2	CAMB
3	0.114	3-4	RAC
	0.105	3-5	CUUI
	0.080	4-5	POST
4	-0.001	6-7	PKU
	-0.022	6-8	UMC
	-0.041	7-10	UFC
	-0.055	8-11	IITB
	-0.062	8-11	INPUT
	-0.074	9-11	SJTU
5	-0.142	12	NTHU
6	-0.358	13	IPN

(c) Human TrueSkill ranking.

	AMU	RAC	CAMB	CUUI	POST	UFC	PKU	UMC	IITB	SJTU	INPUT	NTHU	IPN
AMU	–	.44 ‡	.47 ⋆	.46 †	.44 ‡	.34 ‡	.40 ‡	.37 ‡	.32 ‡	.34 ‡	.32 ‡	.31 ‡	.24 ‡
RAC	**.56** ‡	–	**.53**	.48	.48	.40 ‡	.45 †	.44 ‡	.39 ‡	.38 ‡	.38 ‡	.43 ‡	.28 ‡
CAMB	**.53** ⋆	.47	–	.49	.45 ‡	.43 ‡	.43 ‡	.42 ‡	.42 ‡	.43 ‡	.42 ‡	.43 ‡	.34 ‡
CUUI	**.54** †	**.52**	**.51**	–	.49	.42 ‡	.47	.46 †	.42 ‡	.41 ‡	.41 ‡	.42 ‡	.32 ‡
POST	**.56** ‡	**.52**	**.55** ‡	**.51**	–	.45 ‡	.47	.46 ⋆	.44 ‡	.44 ‡	.43 ‡	.42 ‡	.29 ‡
UFC	**.66** ‡	**.60** ‡	**.57** ‡	**.58** ‡	**.55** ‡	–	**.54** ⋆	**.50**	.49	.44 ⋆	.27 †	.42 ‡	.21 ‡
PKU	**.60** ‡	**.55** †	**.57** ‡	**.53**	**.53**	.46 ⋆	–	.50	.47	.46 ⋆	.46 ⋆	.46 †	.35 ‡
UMC	**.63** ‡	**.56** ‡	**.58** ‡	**.54** †	**.54** ⋆	.50	**.50**	–	.48	.47	.48	.45 ‡	.35 ‡
IITB	**.68** ‡	**.61** ‡	**.58** ‡	**.58** ‡	**.56** ‡	**.51**	**.53**	**.52**	–	.48	.43	.43 ‡	.27 ‡
SJTU	**.66** ‡	**.62** ‡	**.57** ‡	**.59** ‡	**.56** ‡	**.56** ⋆	**.54** ⋆	**.53**	**.52**	–	**.53**	.46 ⋆	.30 ‡
INPUT	**.68** ‡	**.62** ‡	**.58** ‡	**.59** ‡	**.57** ‡	**.73** †	**.54** ⋆	**.52**	**.57**	.47	–	.43 ‡	.22 ‡
NTHU	**.69** ‡	**.57** ‡	**.57** ‡	**.58** ‡	**.58** ‡	**.58** ‡	**.54** †	**.55** ‡	**.57** ‡	**.54** ⋆	**.57** ‡	–	.41 ‡
IPN	**.76** ‡	**.72** ‡	**.66** ‡	**.68** ‡	**.71** ‡	**.79** ‡	**.65** ‡	**.65** ‡	**.73** ‡	**.70** ‡	**.78** ‡	**.59** ‡	–

(d) Head-to-head comparison for ExpectedWins at $p \leq 0.10$ ($\star$), $p \leq 0.05$ ($\dagger$), and $p \leq 0.01$ ($\ddagger$).

Table 3: Comparison of official CoNLL-2014 ranking and human rankings. Ranges and clusters have been calculated with bootstrap resampling at $p \leq 0.05$.

al. (2013) calculate EW scores as follows:

$$\mathrm{score}_{\mathrm{EW}}(S_i) = \frac{1}{|\{S_j\}|} \sum_{j,j \neq i} \frac{\mathrm{win}(S_i, S_j)}{\mathrm{win}(S_i, S_j) + \mathrm{win}(S_j, S_i)}.$$

TrueSkill. The TrueSkill ranking system (Herbrich et al., 2007) is a skill based ranking system for Xbox Live developed at Microsoft Research. It is used to identify and model player (GEC systems in our case) ability in a game to assign players to competitive matches. The TrueSkill ranking system models each player S_i by two parameters: the average relative ability μ_{S_i} and the degree of uncertainty in the player's ability $\sigma^2_{S_i}$. Maintaining uncertainty allows TS to make greater changes to the ability estimates at the beginning and smaller changes after a number of consistent matches has been played. Due to that TS can identify the ability of individual players from a smaller number of pairwise comparisons.

A modification of this approach to the WMT manual evaluation procedure by Sakaguchi et al. (2014) has been adopted as the official ranking method during WMT14 replacing EW. The TrueSkill scores are calculated as inferred means:

$$\mathrm{score}_{TS}(S_i) = \mu_{S_i}.$$

5.2 Rank clusters

Both ranking methods produce total orderings without information on the statistical significance of the obtained ranks. Bojar et al. (2014) notice that the similarity of the participants in terms of methods and training data causes some of them to be very similar and group systems into equivalence classes as proposed by Koehn (2012).

Although the methods and training data among the systems examined in this paper are quite diverse, a great similarity of produced outputs is an inherent property of GEC. Therefore, in this section, for each system S_j placed on rank r_j we also try to determine the true systems rank ranges $[r'_j, \dots, r''_j]$ at a confidence-level of 95% and clusters of equivalent systems by following the procedure outlined by Koehn (2012).

This is accomplished by applying bootstrap resampling. Pairwise rankings are drawn from the set of judgments with multiple drawings. Based on this sample a new ranking is produced. After repeating this process a 1000 times the obtained 1000 ranks for S_j are sorted, with the top 25 and bottom 25 ranks being discarded. The interval of the remaining ranks serves as the final rank range. Next, these rank ranges are used to produce clusters of overlapping rank ranges. This is the last step required to produce the rankings in Tables 3b and 3c for both methods, EW and TS, respectively.

5.3 Choosing the final ranking

Now, we face the question which ranking should be presented as the final result of the human evaluation task. Again, we turn to Bojar et al. (2014) who choose their rankings based on the ranking model's ability to predict pairwise rankings. Accuracy is computed by 100-fold cross-validation. For each fold a new ranking is trained from 99 parts with the left-over part serving as test data.

In a first step, we calculate the accuracy of the unclustered total orderings discarding ties. A ranking based on model scores alone cannot predict ties, this requires equivalence classes. Bojar et al. (2014) define a draw radius r such that systems whose scores differ by less than r are assigned to one cluster, r is tuned to maximize accuracy.

In our case, due to the large number of ties, their method of tuning r is trapped in local maxima and assigns all systems to a single cluster. Alternatively, we propose to calculate clusters according to the method described in the previous section. By fixing $p \leq 0.05$ we directly evaluate rankings of the form given in Table 3. The absolute values of scores and their different interpretations between methods become irrelevant which makes it unnecessary to tune a parameter like r. The main drawback of this approach is its computational cost. For each of the 100 folds we bootstrap another 100 rankings with EW and TS, fix $p \leq 0.05$ and calculate rank clusters. The single clustered ranking for each fold is then used to calculate accuracy for the held-out test data.

Method	EW	TS
Total ordering (non-ties)	58.18	58.15
Bootstrapped clusters	40.12	39.48

Table 4: Accuracy for ranking-based prediction of pairwise judgments.

For our data, contrary to the MT-specific results from Bojar et al. (2014), EW beats TS in both cases (Table 4). We therefore present the ExpectedWins-based ranking (Table 3b) as the final result of the human evaluation effort described in this work and refer to it in the remainder of the paper when the human ranking is mentioned.

5.4 Analysis

The final human-created ranking (Table 3b) consists of four non-overlapping rank clusters. Rank ranges have been calculated at a confidence level of 95%. Comparing the official CoNLL-2014 ranking (Table 3a) with the manually created ExpectedWins ranking shows interesting differences.

The AMU system is judged to be a clear leader by human judges in its own rank cluster. For six out of eight judges, AMU has the highest score (Table 7). The officially winning system CAMB occupies third place in terms of EW scores and is placed in the second cluster with four systems. Only one judge put CAMB in first place. RAC, a middling system, is elevated to second place occupying a rank cluster with three other systems. NTHU, another middling system that based on M^2 should be similar to RAC, is put in the second to last position. Two systems are judged to be worse than INPUT. The rank cluster that includes INPUT is the largest among the four clusters.

We also include pairwise comparisons between all systems according to EW in Table 3d. Each cell contains the percentage of times the system in that column was judged to be better than the system in that row. Bold values mark the winner. We applied the Sign Test to measure statistically significant differences, $\star$ indicates statistical significance at $p \leq 0.10$, † at $p \leq 0.05$, and ‡ at $p \leq 0.01$.

6 Correlation with GEC metrics

Since WMT08 (Callison-Burch et al., 2008) the "metrics task" has been part of the WMT. The aim of the metrics task is to assess the quality of automatic evaluation metrics for MT in terms of correlation with the collected human judgments. We attempt the same in the context of GEC.

6.1 Measures of correlation

Based on Macháček and Bojar (2013), we use Spearman's rank correlation ρ and Pearson's r to compare the similarity of rankings produced by various metrics to the manual ranking from the previous section.

Spearman's rank correlation ρ. Spearman's ρ for rankings with no ties is defined as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the distance between human and metric rank for system i, n is the number of systems.

Pearson's r. Macháček and Bojar (2013) find that Spearman's ρ is too harsh and propose to also use Pearson's r, calculated as

$$r = \frac{\sum_{i=1}^{n}(H_i - \bar{H})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \bar{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \bar{M})^2}}$$

where H and M are the vectors of human and metric scores, $\bar{H}$ and $\bar{M}$ are corresponding means.

6.2 Metrics

The inventory of evaluation metrics for GEC is significantly smaller than for MT. We hope that making our data available will fuel the interest in this area. The following metrics are assessed:

MaxMatch (M^2). Due to its adoption as the main evaluation metric of the CoNLL shared tasks and the QALB shared tasks (Mohit et al., 2014), the M^2 metric (Dahlmeier and Ng, 2012) can be seen as a *de facto* standard. Being an F$_\beta$-score, M^2 results are most influenced by the choice of β. Between the CoNLL-2013 and CoNLL-2014 shared tasks, the organizers changed β from 1.0 to 0.5, and motivate this with intuition alone. The QALB shared tasks for Arabic continue to use $\beta = 1.0$.

Metric	Spearman's ρ	Pearson's r
M^2 F$_{1.0}$	0.648	0.610
M^2 F$_{0.5}$*	0.692	0.627
M^2 F$_{0.25}$	0.720	0.680
M^2 F$_{0.18}$	**0.758**	**0.701**
M^2 F$_{0.1}$	0.670	0.652
I-WAcc	-0.154	-0.098
BLEU	-0.346	-0.240
METEOR	-0.374	-0.241

Table 5: Correlation results for various metrics and human ranking.

	1	2	3	4	5	6	7	8	ρ	$\bar{\rho}$
1	–	.70	.31	.76	.74	.19	.62	.48	.70	.72
2	.72	–	.77	.84	.90	.57	.59	.64	.93	
3	.53	.89	–	.66	.70	.58	.42	.64	.63	
4	.82	.79	.69	–	.91	.42	.67	.54	.91	
5	.65	.85	.82	.87	–	.63	.63	.51	.93	
6	.32	.71	.67	.56	.86	–	.63	.39	.42	
7	.72	.74	.57	.76	.72	.63	–	.63	.76	
8	.64	.85	.86	.69	.72	.57	.75	–	.60	
r	.67	.93	.82	.87	.92	.66	.80	.82	–	
$\bar{r}$	.80									

Table 6: Inter-annotator correlation (Spearman's ρ above the diagonal, Pearson's r below).

I-measure/Weighted Accuracy (I-WAcc). The recently proposed I-WAcc metric (Felice and Briscoe, 2015) tries to address the shortcomings of M^2. The inclusion of true negatives into the formula makes this a very conservative metric; quite similar to the MT metrics described below. The metric assigns negative weights to systems that are harmful with regard to the input text, values from the range $[1, -1]$ are possible. The reported correlation values have been calculated for the ranking presented in Felice and Briscoe (2015).

Machine translation evaluation metrics. Basing most of our results on findings from MT, we also take a look at two machine translation evaluation metrics, BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011). In order to use the CoNLL-2014 gold standard with these metrics, the edit-based annotation has been converted into two plain text files, one per annotator.

6.3 Analysis

The correlation results are collected in table 5. The M^2 metric is generally moderately correlated with

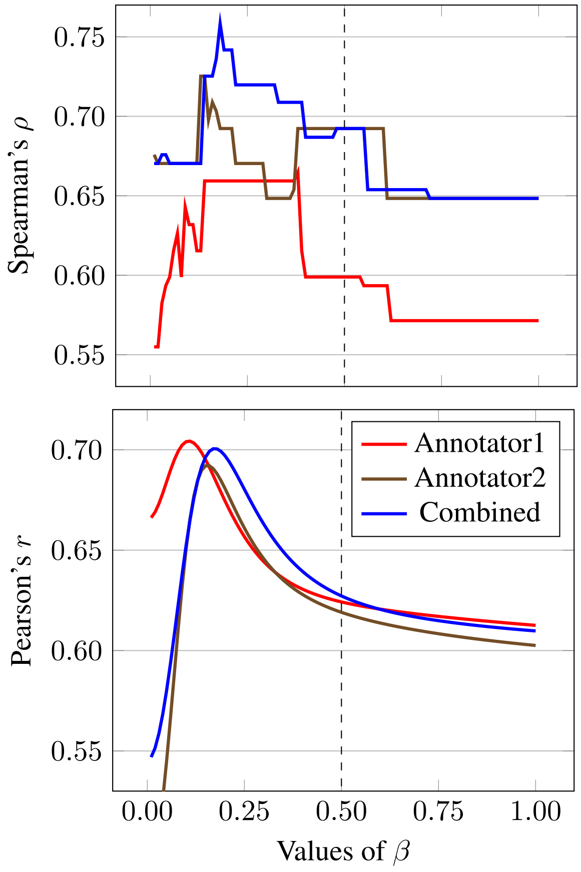


Figure 3: Spearman's ρ and Pearson's r correlation of M^2 with human judgment w.r.t. β. Dashed line marks official CoNLL-2014 choice $\beta = 0.5$.

human judgment and is on the brink of high correlation for values of β closer to 0.2. Compared to M^2, the other metrics are weakly or moderately inversely correlated to human judgment. Inverse correlation with human judgments for metrics that all assign higher scores to better systems seems problematic. In the case of I-WAcc, we would go as far as to state an absence of correlation. It seems the conservative approach adopted for I-WAcc does not correspond to the notion of quality that our judges worked out for themselves. The switch to $\beta = 0.5$ from $\beta = 1.0$ for the CoNLL-2014 shared task was a good choice, but a higher correlation can be achieved for $\beta = 0.25$, the maximum is reached for $\beta = 0.18$. Correlation drops sharply for $\beta = 0.1$. The lack of positive correlation for the MT-metrics is interesting in the light of improvement that results from a shift towards precision for M^2 as BLEU is based on precision.

Figure 3 contains detailed plots of ρ and r with regard to β within the $[0, 1]$ range. As the CoNLL-2014 test data included edits from two annotators, we plot curves for both annotators separately and for the combined gold standard. In the case of Spearman's ρ having alternative error annotations, this leads to higher correlation values. Based on the plots we would recommend setting $0.2 \leq \beta \leq 0.3$ instead of 0.5 or even 1.0.

Inter-annotator correlations of rankings computed for individual judges (Table 6) can be treated as human-level upper bounds for metric correlation. The penultimate column and row contain correlations of rankings for individual judges with rankings computed from all judges minus the respective judge. The last column and row contain the respective weighted (w.r.t. judgments per judge) average of these correlations.

7 Conclusions and future work

We have successfully adapted methods from the WMT human evaluation campaigns to automatic grammatical error correction. The collected and produced data has been made available and should be useful for other researchers. Although we set out to provide answers, we probably ended up with more questions. The following (and more) might be investigated in the future: What makes the winning system special and why do the standard metrics fail at identifying this system? Can we come up with better system-level metrics? Can meaningful sentence-level metrics be developed?

Outside the scope of the particular data, we need to wonder if our results generalize to other shared tasks and other languages. The CoNLL-2014 data concerns ESL learners only and may not be transferable to systems for native speakers. This would be in line with the ideas developed by Chodorow et al. (2012). We would hope to see similar endeavors for the other shared tasks as this would enable the field to draw more general conclusions.

Obtaining the data

The presented data and tools are available from: `https://github.com/grammatical/evaluation`

Acknowledgements

Partially funded by the Polish National Science Centre (Grant No. 2014/15/N/ST6/02330).

The authors would like to thank the following judges for their hard work on the ranking task: Sam Bennett, Peter Dunne, Stacia Levy, Kenneth Turner, and John Winward.

#	Score	Range	System
1	0.674	1-2	CAMB
	0.658	1-2	AMU
2	0.573	3-6	CUUI
	0.573	3-6	PKU
	0.566	3-7	POST
	0.553	3-8	RAC
	0.544	4-8	NTHU
	0.537	5-8	UMC
3	0.436	9-10	SJTU
	0.419	9-11	UFC
	0.387	10-12	IITB
	0.332	11-12	INPUT
4	0.276	13	IPN

(a) Judge 1

#	Score	Range	System
1	0.640	1	AMU
2	0.559	2-6	CUUI
	0.558	2-6	RAC
	0.557	2-7	UFC
	0.543	2-7	CAMB
	0.526	3-9	PKU
	0.511	5-10	POST
	0.511	4-11	IITB
	0.508	5-10	UMC
	0.473	7-11	INPUT
	0.472	8-11	NTHU
3	0.398	12	SJTU
4	0.286	13	IPN

(b) Judge 2

#	Score	Range	System
1	0.612	1-2	AMU
	0.578	2-3	CUUI
	0.564	2-5	UFC
	0.534	3-7	RAC
	0.526	4-7	POST
	0.517	4-8	UMC
	0.508	4-9	IITB
	0.474	5-11	INPUT
	0.467	8-12	SJTU
	0.464	8-12	PKU
	0.463	8-12	CAMB
	0.458	9-12	NTHU
2	0.333	13	IPN

(c) Judge 3

#	Score	Range	System
1	0.641	1-2	AMU
	0.631	1-2	CAMB
2	0.581	3-4	RAC
	0.578	3-4	CUUI
3	0.507	5-10	UFC
	0.494	5-10	POST
	0.490	5-10	UMC
	0.488	5-11	SJTU
	0.486	5-10	PKU
	0.473	5-11	INPUT
	0.441	9-12	NTHU
	0.378	11-12	IITB
4	0.313	13	IPN

(d) Judge 4

#	Score	Range	System
1	0.613	1-2	AMU
	0.608	1-2	RAC
2	0.568	3-5	CUUI
	0.554	3-6	CAMB
	0.535	4-7	POST
	0.526	4-9	UFC
	0.515	5-9	PKU
	0.496	6-10	SJTU
	0.487	6-11	INPUT
	0.472	8-11	IITB
	0.461	9-11	UMC
3	0.375	12	NTHU
4	0.290	13	IPN

(e) Judge 5

#	Score	Range	System
1	0.601	1-2	RAC
	0.579	1-4	AMU
	0.565	1-6	IITB
	0.562	2-6	POST
	0.548	2-8	INPUT
	0.535	3-8	UFC
	0.519	6-9	PKU
	0.516	6-9	CAMB
	0.491	7-10	SJTU
	0.474	9-10	CUUI
2	0.428	11	UMC
3	0.368	12	NTHU
4	0.313	13	IPN

(f) Judge 6

#	Score	Range	System
1	0.788	1	AMU
2	0.697	2	CAMB
3	0.553	3-7	RAC
	0.544	3-7	UMC
	0.537	3-7	POST
	0.533	3-10	IITB
	0.487	5-10	PKU
	0.474	5-12	INPUT
	0.462	6-11	CUUI
	0.436	6-12	SJTU
	0.408	8-12	UFC
	0.386	9-12	NTHU
4	0.303	13	IPN

(g) Judge 7

#	Score	Range	System
1	0.660	1-2	AMU
	0.625	1-3	CUUI
	0.568	2-7	IITB
	0.556	3-6	CAMB
	0.543	3-7	UMC
	0.537	3-7	POST
	0.483	5-10	INPUT
	0.481	5-10	UFC
	0.478	7-11	RAC
	0.466	7-11	NTHU
	0.420	10-12	PKU
	0.419	10-12	SJTU
2	0.260	13	IPN

(h) Judge 8

Table 7: Rankings by individual annotators. Cluster ranks and rank ranges have been computed with bootstrap resampling at $p \leq 0.1$ to accomodate for the reduced number of judgments per judge.

References

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 1–44. ACL.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 12–58. ACL.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc. of the Third Workshop on Statistical Machine Translation*, pages 70–106. ACL.

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

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

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proc. of the 2012 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 568–572. ACL.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proc. of the 13th European Workshop on Natural Language Generation*, pages 242–249. ACL.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. Hoo 2012: A report on the preposition and determiner error correction shared task. In *Proc. of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62. ACL.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proc. of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Christian Federmann. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. ELRA.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proc. of the 2015 Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL-HLT 2015)*. ACL.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. In *International Workshop on Spoken Language Translation*, pages 179–184.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 45–51. ACL.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 293–301. ACL.

Nitin Madnani, Joel Tetreault, Martin Chodorow, and Alla Rozovskaya. 2011. They can help: using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proc. of the 49th Annual Meeting of the ACL: Human Language Technologies*, pages 508–513. ACL.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for arabic. *Proc. of the EMNLP 2014 Workshop on Arabic Natural Language Processing*, pages 39–47.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammaticality correction metrics. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 588–593. ACL.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proc. of the 17th Conference on Computational Natural Language Learning*, pages 1–12. 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 *Proc. of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–14. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting on ACL*, pages 311–318. ACL.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 1–11. ACL.