Evaluating the performance of Automated Text Scoring systems

Helen Yannakoudakis
The ALTA Institute
Computer Laboratory
University of Cambridge
helen.yannakoudakis@cl.cam.ac.uk

Ronan Cummins
The ALTA Institute
Computer Laboratory
University of Cambridge
ronan.cummins@cl.cam.ac.uk

Abstract

Various measures have been used to evaluate the effectiveness of automated text scoring (ATS) systems with respect to a human gold standard. However, there is no systematic study comparing the efficacy of these metrics under different experimental conditions. In this paper we first argue that *measures of agreement* are more appropriate than *measures of association* (i.e., correlation) for measuring the effectiveness of ATS systems. We then present a thorough review and analysis of frequently used *measures of agreement*. We outline desirable properties for measuring the effectiveness of an ATS system, and experimentally demonstrate using both synthetic and real ATS data, that some commonly used measures (e.g., Cohen's kappa) lack these properties. Finally, we identify the most appropriate *measures of agreement* and present general recommendations for best evaluation practices.

1 Introduction

Automated assessment of text was introduced in the early 1960s in an attempt to address several issues with manual assessment (e.g., expense, speed, and consistency). Further advantages become more pronounced when it comes to scoring extended texts such as essays, a task prone to an element of subjectivity. Automated systems enable rigid application of scoring criteria, thus reducing the inconsistencies which may arise, in particular, when many human examiners are employed for large-scale assessment.

There is a substantial literature describing and evaluating ATS systems (Page, 1968; Powers et al., 2002; Rudner and Liang, 2002; Burstein et al., 2003; Landauer et al., 2003; Higgins et al., 2004; Attali and Burstein, 2006; Attali et al., 2008; Williamson, 2009; Briscoe et al., 2010; Chen and He, 2013). Such systems are increasingly used but remain controversial. Although a comprehensive comparison of the capabilities of eight existing commercial essay scoring systems (Shermis and Hamner, 2012) across five different performance metrics in the recent ATS competition organised by Kaggle[1] claimed that ATS systems grade similarly to humans, critics (Wang and Brown, 2007; Wang and Brown, 2008; Perelman, 2013) have continued to dispute this.

For the evaluation of ATS systems (Williamson, 2009; Williamson et al., 2012), emphasis has been given to the "agreement" of machine-predicted scores (ordinal grades) with that of a human gold standard; that is, scores assigned by human examiners to the same texts that the machine is evaluated on. Various metrics have been used, the most prominent being Pearson's correlation, percentage of agreement, and variations of Cohen's kappa statistic. Inconsistencies in the reporting of, and misconceptions in the interpretation of, these metrics in published work makes cross-system comparisons on publicly-available datasets more difficult. The lack of careful motivation of any metric fuels opposition to the deployment of ATS. To date, several ATS systems are being used operationally for high-stakes assessment in addition to them being part of self-assessment and self-tutoring sys-

[1] https://www.kaggle.com/c/asap-aes

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 213–223,
Denver, Colorado, June 4, 2015.

tems, underscoring the need for common and well-motivated metrics that establish true system performance.

In this paper, we define the task of ATS as the accurate prediction of gold-standard scores (predefined ordinal grades), and we experimentally examine the robustness and efficacy of *measures of agreement* for a number of different conditions under two different experimental setups. First, we use synthetic data to simulate various experimental conditions, and second, we use real ATS data to assess the effectiveness of the metrics under realistic scenarios. For the latter, we run a series of experiments on the output of state-of-the-art ATS systems. We outline some deficiencies in commonly used metrics that have been previously overlooked, and consequently we propose more appropriate metrics for evaluating ATS systems focusing primarily on optimising system effectiveness and facilitating cross-system comparison.

The focus on measures of agreement is motivated by their use as the primary metric for evaluating system effectiveness in the recent Kaggle essay scoring competition. To the best of our knowledge, there is no systematic study comparing the efficacy of different measures of agreement under different experimental conditions. Although we focus on the task of ATS, the recommendations regarding the metrics covered in this paper extend naturally to many similar NLP tasks, i.e., those where the task is to accurately predict a gold-standard score.

The remainder of the paper is structured as follows: Section 2 defines our task and objectives. Section 3 reviews a number of performance metrics relevant to the ATS task. Section 4 describes a set of desired metric properties and presents an analysis of some prominently used metrics for the ATS task that uses the output of both simulated and real systems. Section 5 concludes with a discussion, general recommendations for evaluation practices and an outline of future work.

2 Task Definition

In the standard ATS evaluation, there exists a set of n texts where each text is indexed t_1 to t_n. Each text t_i is assigned a gold standard score $gs(t_i)$ by a human assessor (or group of human assessors). This score is one of g ordinal scores, which for convenience we index 1 to g. It is worth noting that, in general, it is not a requirement that the differences in scores are uniform. Furthermore, there exists some number of ATS systems ats_j indexed $j = 1$ to $j = m$ that predict scores $ats_j(t_i)$ for each of the n texts.

Given two ATS systems ats_1 and ats_2, we would like to determine a metric $\mathcal{M}$ that returns a measure of performance for ats_1 and ats_2 for which $\mathcal{M}(ats_1, gs, t) > \mathcal{M}(ats_2, gs, t)$ when ats_1 is a *better* system than ats_2. Note that we have not defined what "better" means at this stage. We will return to describing some desirable properties of $\mathcal{M}$ in Section 4.

From an educational point of view, our task is to ascertain whether the writing abilities required to warrant a particular score/grade have been attained. From this perspective, measures of agreement seem the appropriate type of measurement to apply to the output of ATS systems to address the accuracy of the (numerical) solution compared to the gold standard.

3 Measuring Performance of ATS systems

In this section, we review and critique metrics that have been frequently used in the literature to ascertain the performance of ATS systems. These performance metrics can be broadly categorised into *measures of association* and *measures of agreement* (e.g., see Williamson et al. (2012)).

3.1 Measures of Association

Measures of association (i.e., correlation coefficients) have been widely used in ATS (e.g., Yannakoudakis et al. (2011)), with Pearson's product-moment correlation coefficient being the most common. Pearson's correlation is a parametric measure of association that quantifies the degree of linear dependence between two variables and, more specifically, describes the extent to which the variables co-vary relative to the degree to which they vary independently. The greater the association, the more accurately one can use the value of one variable to predict the other. As the data depart from the coefficient's assumptions (e.g., unequal marginals), its maximum values may not be attainable (Carroll, 1961). For ordinal data, unequal marginals will al-

ways involve ties.[2] As the number of ties increases relative to the number of observations, its appropriateness largely diminishes.[3]

Spearman's rank correlation coefficient is a non-parametric measure of association that has the same range as Pearson, and it is calculated by ranking the variables and computing Pearson on the ranks rather than the raw values. In contrast to Pearson, it assesses the strength of a monotonic rather than linear relation between two variables, and has the advantage of independence from various assumptions. Unlike Pearson, it exhibits robustness to outliers; however, its reliability also decreases as the number of ties increases. It is worth noting at this point Kendall's τ_b, which is a more effective tie-adjusted non-parametric bivariate coefficient that quantifies the degree of agreement between rankings, and it is defined in terms of concordant and discordant pairs, although ties also affect its reliability.

3.1.1 Discussion

In essence, non-parametric measures are measures of *rank* correlation. In the context of the evaluation of ATS, they measure agreement with respect to the ranks, that is, whether an ATS system ranks texts similarly to the gold standard. However, this is not an appropriate type of measurement given the task definition in Section 2, where we would like to ascertain actual agreement with respect to the scores. Furthermore, correlation coefficients do not account for any systematic biases in the data; for example, a high correlation can be observed even if the predicted scores are consistently n points higher than the gold standard.

In the presence of outliers, the coefficient can be misleading and pulled in either direction. For example, for Pearson's correlation an outlier can influence the value of the correlation to the extent that a high correlation is observed even though the data may not be linearly dependent. Furthermore, it is well known that the value of the correlation will be greater if there is more variability in the data than if there is less. This is caused by the mathematical constraints in their formulation, and does not necessarily reflect the true relationship of predicted to gold standard scores. Finally, their reliability decreases as the number of ties increases. We come back to the appropriateness and recommended use of correlation metrics in Section 5.

In summary, (non-parametric) correlation measures are more apt at measuring the ability of the ATS system to correctly *rank* texts (i.e., placing a well written text above a poorly written text), rather than the ability of the ATS system to correctly assign a score/grade. In other words, correlation measures do not reward ATS systems for their ability to correctly identify the thresholds that separate score/grade boundaries ($1:2$ to $g-1:g$).

3.2 Measures of Agreement

A simple way of gauging the agreement between gold and predicted scores is to use percentage agreement, calculated as the number of times the gold and predicted scores are the same, divided by the total number of texts assessed. A closely-related variant is percentage of adjacent agreement, in which agreement is based on the number of times the gold and predicted scores are no more than n points apart.

Despite its simplicity, it has been argued (Cohen, 1960) that this measure can be misleading as it does not exclude the percentage of agreement that is expected on the basis of pure chance. That is, a certain amount of agreement can occur even if there is no systematic tendency for the gold and predicted scores to agree. The kappa coefficient (C_κ) was introduced by Cohen (1960) as a measure of agreement adjusted for chance. Let P_a denote the percentage of observed agreement and P_e the percentage of agreement expected by chance, Cohen's kappa coefficient is calculated as the ratio between the "true" observed agreement and its maximum value:

$$C_\kappa = \frac{P_a - P_e(\kappa)}{1 - P_e(\kappa)} \quad (1)$$

where $P_e(\kappa)$ is the estimated agreement due to chance, and is calculated as the inner-product of the marginal distribution of each assessor (a worked example of this follows in Section 3.2.1). The values of the coefficient range between -1 and 1, where 1 represents perfect agreement and 0 represents no agreement beyond that occurring by chance. Most

[2] Of course, ties exist even when the marginals are identical if the number of observations is larger than the number of scores.

[3] For more details see (Maimon et al., 1986; Goodwin and Leech, 2006; Hauke and Kossowski, 2011) among others.

measures of agreement that are corrected for chance agreement, of which there are many, follow the general formula above where P_e varies depending on the specific measure. The disadvantage of this basic measure applied to ordinal data (scores in our case) is that it does not allow for weighting of different degrees of disagreement.

Weighted kappa (Cohen, 1968) was developed to address this problem. Note that this was the main metric used for evaluation and cross-system comparison of essay scoring systems in the recent Kaggle shared-task competition on ATS. It is commonly employed with ordinal data and can be defined either in terms of agreement or disagreement weights. The most common weights used are the (absolute) linear error weights and the quadratic error weights (Fleiss, 1981). The linear error weights are proportional to the actual difference between the predicted scores and the gold standard, while the quadratic error weights are proportional to the squared actual difference between these scores. The choice of weights is important as they can have a large effect on the results (Graham and Jackson, 1993).

In what follows, we discuss two of the main problems regarding the kappa coefficient: its dependency on trait prevalence and on marginal homogeneity. We note that the properties kappa exhibits (as shown below) are *independent* of the type of data on which it is used, that is, whether there is a categorical or an ordinal (gold standard) scale.

3.2.1 Trait Prevalence

Trait prevalence occurs when the underlying characteristic being measured is not distributed uniformly across items. It is usually the case that gold standard scores are normally distributed in the ATS task (i.e., the scores/grades are biased towards the mean).

Table 1 shows an example of the effect of trait prevalence on the C_κ statistic using a contingency table. In this simple example there are two scores/grades (i.e., pass P or fail F) for two different sets of 100 essays with different gold-score distributions, gs_1 and gs_2. The rows of the matrix indicate the frequency of the scores predicted by the ATS, while the columns are the gold-standard scores. Although percentage agreement (along the main diagonal) in both cases is quite high, $P_a = 0.8$, the C_κ statistic varies quite considerably. As the observed marginals (i.e., the totals either vertically or horizontally, or otherwise, the distribution of the scores / grades) in ats$\backslash$gs$_1$ are uniformly distributed, the correction for chance agreement is much lower (i.e., $P_e(\kappa) = 0.5\times0.5+0.5\times0.5 = 0.5$) than for ats$\backslash$gs$_2$ (i.e., $P_e(\kappa) = 0.68 \times 0.8 + 0.32 \times 0.2 = 0.61$) with unequal marginals, which leads to a lower absolute C_κ value for ats$\backslash$gs$_2$. In this example, it is not clear why one would want a measure of agreement with this behaviour, where P_e is essentially artificially increased when the marginals are unequal.

ats $\backslash$gs$_1$	P	F	
P	40	10	50
F	10	40	50
	50	50	100

ats $\backslash$gs$_2$	P	F	
P	64	4	68
F	16	16	32
	80	20	100

Table 1: Cohen's κ for an ats system on two sets of essays. Although percentage agreement is 0.8 for both sets of essays, $C_\kappa = 0.6$ (left) and $C_\kappa = 0.49$ (right).

Fundamentally, this implies that the comparison of systems across datasets (or indeed the comparison of datasets given the same system) is very difficult because the value of C_κ not only depends on actual agreement, but crucially also on the distribution of the gold standard scores.

3.2.2 Marginal Homogeneity

A second problem with C_κ is that the difference in the marginal probabilities affects the coefficient considerably. Consider Table 2, which shows two different ATS system ratings (ats_1 and ats_2) along the same gold standard scores. The value of C_κ for ats_2 is much smaller (and actually it is 0) compared to that for ats_1 (0.12), even though ats_2 has higher percentage and marginal agreement; that is, ats_2 predicts scores with frequencies that are more similar to those in the gold standard.[4] The root cause of this paradox is similar to the one described earlier, and arises from the way P_e is calculated, and more specifically the assumption that marginal probabilities are classification propensities that are fixed, that is, they are known to the assessor before classifying the instances/texts into categories/scores. This

[4]Of course higher marginal agreement does not translate to overall higher agreement if percent agreement is low.

ats_1 \gs	P	F	
P	20	0	20
F	60	20	80
	80	20	100

ats_2 \gs	P	F	
P	60	15	75
F	20	5	25
	80	20	100

Table 2: C_κ for two systems ats_1 and ats_2 for the same set of gold scores. Although percentage agreement for ats_1 and ats_2 is 0.4 and 0.65 respectively, C_κ for ats_1 and ats_2 is $C_\kappa = 0.12$ (left) and $C_\kappa = 0$ (right).

is clearly not the case for ATS systems, and therefore the dependence of chance agreement on the level of marginal agreement is questionable when the marginals are free to vary (Brennan and Prediger, 1981).[5]

Essentially, the end result when a system predicts scores with a marginal distribution that is more similar to the gold standard (i.e., ats_2), is that any misclassification is penalised more severely even though percent agreement may be high. This is not the behaviour we want in a performance metric for ATS systems. Using kappa as an objective function in any machine learning algorithm could easily lead to learning functions that favour assigning distributions that are *different* to that of the gold standard (e.g., Chen and He (2013)).

3.2.3 Discussion

Previous work has also demonstrated that there exist cases where high values of (quadratic) kappa can be achieved even when there is low agreement (Graham and Jackson, 1993). Additionally, Brenner and Kliebsch (1996) investigated the effect of the score range on the magnitude of weighted kappa and found that the quadratic weighted kappa coefficient tends to have high variation and increases as the score range increases, particularly in ranges between two and five distinct scores. In contrast, linearly weighted kappa appeared to be less affected, although a slight increase in value was observed as the range increased.

The correction for chance agreement in Cohen's kappa has been the subject of much controversy (Brennan and Prediger, 1981; Feinstein and Cicchetti, 1990; Uebersax, 1987; Byrt et al., 1993; Gwet, 2002; Di Eugenio and Glass, 2004; Sim and Wright, 2005; Craggs and Wood, 2005; Artstein and Poesio, 2008; Powers, 2012). Firstly, it assumes that when assessors are unsure of a score, they *guess* at random according to a fixed prior distribution of scores. Secondly, it includes chance correction for every single prediction instance (i.e., not only when an assessor is in doubt). Many have argued (Brennan and Prediger, 1981; Uebersax, 1987) that this is a highly improbable model of assessor error and vastly over-estimates agreement due to chance, especially in the case when prior distributions are free to vary. Although it is likely that there is some agreement due to chance when an assessor is unsure of a score (Gwet, 2002), it is unlikely that human assessors simply guess at random, and it is unlikely that this happens for *all* predictions. For the task of ATS, the distribution of scores to assign are not *fixed* a priori. Although trained assessors may have a certain expectation of the final distribution, it is certainly not fixed.[6]

Consequently, there are a number of different agreement metrics – for example, Scott's π (Scott, 1955), which is sensitive to trait prevalence but not the marginals, and Krippendorff's α (Krippendorff, 1970) which is nearly equivalent to π (Artstein and Poesio, 2008) – all of which vary in the manner in which chance agreement (i.e., P_e) is calculated, but have similar problems (Zhao, 2011; Gwet, 2014). It is also worth noting that weighted versions of kappa do not solve the issues of trait prevalence and marginal homogeneity.

The two most noteworthy variants are the *agreement coefficient* AC (Gwet, 2002) and the Brennan-Prediger (BP) coefficient (Brennan and Prediger, 1981), which both estimate P_e more conservatively using more plausible assumptions. In particular, the BP coefficient estimates P_e using $1/g$, with the assumption that the probability that an assessor would guess the score of an item by chance is inversely related to the number of scores g in the rating scale.[7] Substituting P_e in equation (1) gives $(P_a - 1/g)/(1 - 1/g)$, which is better suited when one or both of the marginals are free to vary. When

[5] However, we would like to penalise trivial systems that e.g., always assign the most prevalent gold score, in which case the marginals are indeed fixed.

[6] See Brennan and Prediger (1981) for a more detailed discussion.

[7] We note that this is equivalent to the S coefficient (Bennett et al., 1954) discussed in (Artstein and Poesio, 2008).

the grades are not uniformly distributed, P_e may be higher than $1/g$; nevertheless, it can be a useful lower limit for P_e (Lawlis and Lu, 1972). Note that in the example in Table 1, BP would be the same for both ats\gs$_1$ and ats\gs$_2$, $(0.8 - 0.5)/(1 - 0.5) = 0.6$, and thus effectively remains unaltered under the effects of trait prevalence.[8]

The AC coefficient calculates P_e as follows:

$$P_e = \frac{1}{(g-1)} \sum_{k=1}^{g} \pi_k(1 - \pi_k) \quad (2)$$

$$\pi_k = (p_{a,k} + p_{b,k})/2 \quad (3)$$

where π_k represents the probability of assigning grade/score k to a randomly selected item by a randomly selected assessor, calculated based on $p_{a,k}$ and $p_{b,k}$, which are the marginal probabilities of each assessor a and b respectively for grade/score k. More specifically, $p_{a,k} = n_{a,k}/n$ and $p_{b,k} = n_{b,k}/n$, where $n_{a,k}$ refers to the number of instances assigned to grade k by assessor a, $n_{b,k}$ refers to the number of instances assigned to grade k by assessor b, and n refers to the total number of instances. Gwet (2002;2014) defines chance agreement as the product of the probability that two assessors agree given a non-deterministic instance,[9] defined as $1/g$, by the propensity for an assessor to assign a non-deterministic grade/score, estimated as $\sum_{k=1}^{g} \pi_k(1 - \pi_k)/(1 - 1/g)$.[10]

In the example in Table 1, $P_e = (0.5 \times (1-0.5) + 0.5 \times (1 - 0.5))/(2 - 1) = 0.5$ for ats\gs$_1$ (for which $\pi_{pass} = \pi_{fail} = 0.5$), and $P_e = (0.74 \times (1 - 0.74) + 0.26 \times (1-0.26))/(2-1) = 0.38$ for ats\gs$_2$, which is in contrast to C_κ that overestimated P_e for ats\gs$_2$ with unequal marginals. More specifically, the AC coefficient would be higher for ats\gs$_2$ than for ats\gs$_1$: 0.67 versus 0.60 respectively.[11]

4 Metric Properties

On the basis of the discussion so far, we propose the following list of desirable properties of an evaluation measure for an ATS system:

- Robustness to trait prevalence
- Robustness to marginal homogeneity
- Sensitivity to magnitude of misclassification
- Robustness to score range

In this section, we analyse the aforementioned metrics of agreement (with different weights) with respect to these properties using both synthetic and real ATS-system scores (where applicable).

4.1 Robustness to Trait Prevalence

In order to test metrics for robustness to trait prevalence, we simulated 5,000 gold standard scores on a 5-point scale using a Gaussian (normal) distribution with a mean score at the mid-point. By controlling the variance of this Gaussian, we can create gold standard scores that are more *peaked* at the center (high trait prevalence) or more uniform across all grades (low trait prevalence). We simulated systems by randomly introducing errors in these scores. The system output in Figure 1 (left) was created by randomly sampling 25% of the gold standard scores and perturbing them by 2 points in a random direction.[12] This led to a simulated system with 75% percentage agreement, which also translates to a constant mean absolute error (MAE) of 0.5 (i.e., on average, each predicted score is 0.5 scores away from its gold counterpart).

Figure 1 (left) shows that nearly all evaluation measures are very sensitive to the distribution of the gold standard scores, and the magnitude of the metrics does change as the distribution becomes more peaked. The AC measure is less sensitive than C_κ (and actually rewards systems), but the only measure of agreement that is invariant under changes in trait prevalence is the BP coefficient, which actually is in line with percentage agreement and assigns 75% agreement using quadratic weights.

To study the effect of trait prevalence on real systems, we replicated an existing state-of-the-art ATS system (Yannakoudakis et al., 2011). The model

[8] However, it can be artificially increased as the scoring scale increases.

[9] That is, it is a hard-to-score instance, which is the case where random ratings occur.

[10] See (Gwet, 2014) for more details regarding the differences between AC and Aickin's alpha (Aickin, 1990).

[11] The reader is referred to (Gwet, 2014; Brennan and Prediger, 1981) for more details on AC and BP and their extensions to at least ordinal data and to more than two assessors.

[12] If this could not be done (i.e., a score of 4 cannot be changed by +2 on a 5-point scale), a different score was randomly sampled.

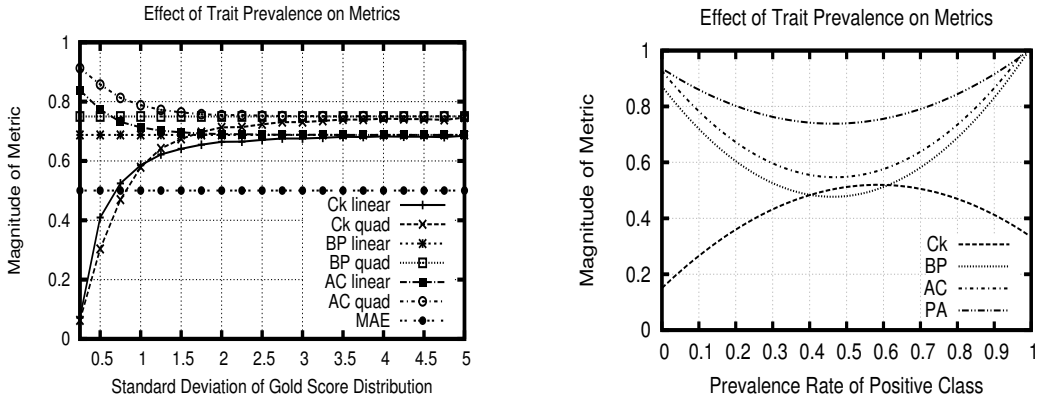

Figure 1: Effect of trait prevalence on metrics of agreement for synthetic (left) and real (right) ATS scores.

evaluates writing competence on the basis of lexical and grammatical features, as well as errors, and achieves a correlation of around 0.75 on the publicly available First Certificate in English (FCE) examination scripts, that have been manually annotated with a score in the range 1 to 40 (with 40 being the highest). In this setting, robustness to trait prevalence was evaluated by plotting the magnitude of the metrics as a function of the prevalence rate, calculated as the proportion of passing scores in the data, as judged by both the ATS system and the examiner, as we varied the passing threshold from 1 to 40.

In Figure 1 (right) we can see that all metrics are sensitive to trait prevalence.[13] In order to get a clearer picture on the effect of trait prevalence on real systems, it is important we plot percent agreement (PA) along with the metrics. The reason is that chance-corrected agreement measures should remain reasonably close to the quantity that they adjust for chance, as this quantity varies (Gwet, 2014). AC and BP remain reasonably close to PA as the prevalence rate increases. On the other hand, C_κ is further away, and at times considerably lower. The behaviour of kappa is difficult to explain, and in fact, even when the prevalence rate approaches 1, C_κ still produces very low results. Note that in the binary pass/fail case, linear and quadratic weights do not affect the value of kappa and produce the same results.

[13]Curve fitting is performed to be able to observe the tendency of the metrics.

4.2 Robustness to Marginal Homogeneity

In order to test the metrics for robustness to marginal homogeneity, we simulated 5,000 gold standard scores on a 10-point scale using a Gaussian distribution with a mean score at the mid-point and a standard deviation of one score. We simulated different systems by randomly introducing errors in these scores. In particular, we simulated outputs that had distributions different to that of the gold standard by drawing a number of incorrect scores from a different Gaussian centred around a different mean (0–9 in Figure 2). We kept percentage agreement with linear weights constant, which again also translates to a constant MAE (1.0). We are looking for metrics that are less sensitive to varying marginals, and ideally which promote systems that distribute scores similarly to the gold standard when agreement is otherwise identical.

For the measures of agreement, as expected, Cohen's kappa (both linear and quadratic) penalises systems that distribute scores similarly to those of the gold standard. However, AC (with linear and quadratic weights) and quadratic BP promote systems that distribute scores similarly to the gold scores. On the other hand, BP linear remains unchanged.

To study the sensitivity of the metrics to the variations in the marginal distributions in real ATS systems, we plot their variation as a function of the similarity of the passing-score distributions, where the similarity is calculated as $sim_{pass} = 1 - |p_{gold,pass} - p_{ats,pass}|$, which is based on the absolute difference

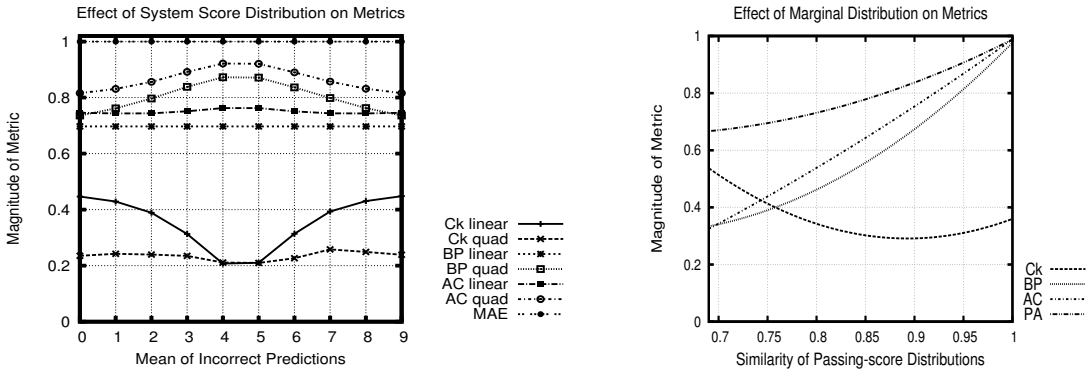

Figure 2: Sensitivity of metrics on the marginal distribution of synthetic (left) and real (right) ATS scores.

between the marginal probabilities of the gold and predicted passing scores. The higher the value of sim, the more similar the distributions are. Again, we employ Yannakoudakis et al. (2011)'s ATS system.

Similarly to the simulated experiments, we observe increases in the magnitude of AC and BP as the similarity increases, whereas C_κ is considerably lower and has a decreasing tendency which does not stay close to PA. In fact, marginal homogeneity does not guarantee the validity of the results for Cohen's kappa. This can be seen more clearly in Figure 3, where we plot the magnitude of the metrics as a function of the overall probability of assigning a passing score, as judged by both the human assessor and the ATS system. That is, $(p_{gold,pass} + p_{ats,pass})/2$. As the overall probability of a passing score becomes very large or very small, C_κ yields considerable lower results, regardless of whether the marginal probabilities are equal or not.

4.3 Sensitivity to Magnitude of Misclassification

It is common that human assessors disagree by small margins given the subjectivity of the ATS task. However, larger disagreements are usually treated more seriously. Therefore, given two ATS systems, we would prefer a system that makes more small misclassifications over a system that makes a few large misclassifications when all else is equal. A metric with quadratic-weighting is likely to adhere to this property.

To test the sensitivity of the metrics to the magnitude of misclassification, we simulated 5,000 gold standard scores on a 10-point scale using a Gaussian (normal) distribution with a mean score at the midpoint. Again, we simulated systems by randomly introducing errors to the scores. For each system, we varied the magnitude of the misclassification while the total misclassification distance (i.e., MAE or PA) was kept constant. Figure 4 confirms that measures of agreement that use quadratic weights decrease as the magnitude of each error increases. The metrics of agreement that use linear weights actually increase slightly.[14]

4.4 Robustness to score scales

Robustness of the metrics to the score range or scale was tested by binning the gold and predicted scores at fixed cutpoints and re-evaluating the results. In the FCE dataset, the scale was varied between 40 and 3 points by successively binning scores. Metrics that are less sensitive to scoring scales facilitate cross-dataset comparisons.

All metrics were affected by the scale, although those with quadratic weights appeared to be more sensitive compared to those with linear ones. Quadratic C_κ was the most sensitive metric and showed larger decreases compared to the others as the scoring scale was reduced, while AC quadratic exhibited higher stability compared to BP quadratic.[15]

[14] Note that such an experiment cannot be controlled and reliably executed for real systems.

[15] Detailed results omitted due to space restrictions.

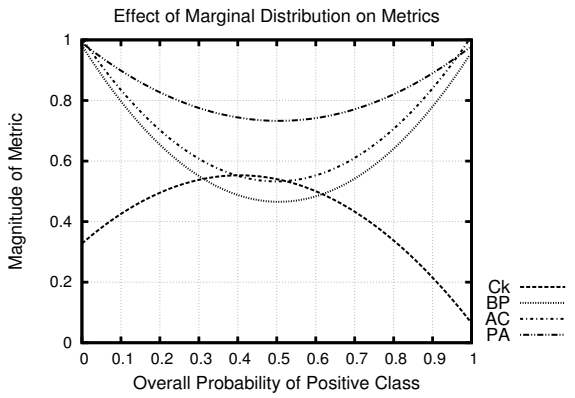

Figure 3: Sensitivity of metrics on the marginal distribution of real ATS-model scores.

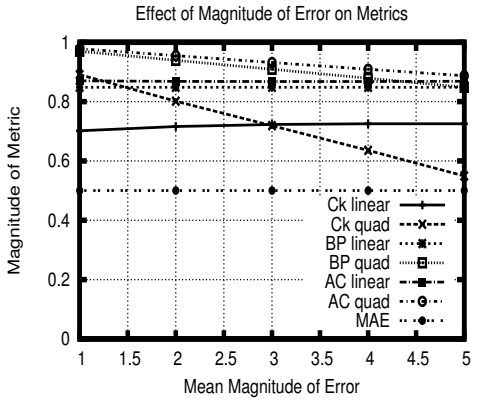

Figure 4: Change in the magnitude of performance metrics as only the size of each misclassification increases.

5 Recommendations and Conclusion

Our results suggest that AC and BP (with quadratic weights) overall are the most robust agreement coefficients. On the basis of this analysis, we make the following recommendations:

- We recommend against using Cohen's kappa. Interpretation of the magnitude of kappa within / across system and dataset comparisons is problematic, as it depends on trait prevalence, the marginal distributions and the scoring scale. It is worth noting at this point that the inefficacy of kappa is *independent* on the type of data that it is being used on, that is, whether there is a categorical or ordinal (gold standard) scale.

- We recommend using the AC coefficient with quadratic weights. Although BP is a good alternative as it adjusts percent agreement simply based on the inverse of the scoring scale, it is more sensitive to, and directly affected by the scoring scale.

- We recommend *reporting* a rank correlation coefficient (Spearman's or Kendall's τ_b rank correlation coefficient), rather than using it for system evaluation and comparison, as it can facilitate error analysis and system interpretation; for example, low agreement and high rank correlation would indicate a large misclassification magnitude, but high agreement with respect to the ranking (i.e., the system ranks texts similarly to the gold standard); high agreement and low rank correlation would indicate high accuracy in predicting the gold scores, but small ranking errors.[16] Kendall's τ_b may be preferred, as it is a more effective tie-adjusted coefficient that is defined in terms of concordant and discordant pairs; however, further experiments beyond the scope of this paper would be needed to confirm this.

It is worth noting that given the generality of the ATS task setting as presented in this paper (i.e., aiming to predict gold standard scores on an ordinal scale) and the metric-evaluation setup (using synthetic data in addition to real output), the properties discussed and resulting recommendations may be more widely relevant within NLP and may serve as a useful benchmark for the wider community (Siddharthan and Katsos, 2010; Bloodgood and Grothendieck, 2013; Chen and He, 2013; Liu et al., 2013, among others) as well as for shared task organisers.

An interesting direction for future work would be to explore the use of evaluation measures that lie outside of those commonly used by the ATS community, such as macro-averaged root mean squared error that has been argued as being suitable for ordinal regression tasks (Baccianella et al., 2009).

[16] A low correlation could also point to effects of the underlying properties of the data as the metric is sensitive to trait prevalence (see Section 3.1.1).

Acknowledgments

We would like to thank Ted Briscoe for his valuable comments and suggestions, Cambridge English Language Assessment for supporting this research, and the anonymous reviewers for their useful feedback.

References

Mikel Aickin. 1990. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*, 46(2):293–302.

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

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-Rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3):1–30.

Yigal Attali, Don Powers, Marshall Freedman, Marissa Harrison, and Susan Obetz. 2008. Automated Scoring of short-answer open-ended GRE subject test items. Technical Report 04, ETS.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *9th IEEE International Conference on Intelligent Systems Design and Applications*, pages 283–287. IEEE Comput. Soc.

E M Bennett, R Alpert, and A C Goldstein. 1954. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308.

Michael Bloodgood and John Grothendieck. 2013. Analysis of Stopping Active Learning based on Stabilizing Predictions. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 10–19.

Robert L Brennan and Dale J Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699.

Hermann Brenner and Ulrike Kliebsch. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 7(2):199–202.

Ted Briscoe, Ben Medlock, and Øistein E. Andersen. 2010. Automated assessment of ESOL free text examinations. Technical Report UCAM-CL-TR-790, University of Cambridge, Computer Laboratory, nov.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003. Criterion: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence*, pages 3–10.

Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429.

John B. Carroll. 1961. The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 26(4):347–372, December.

Hongbo Chen and Ben He. 2013. Automated Essay Scoring by Maximizing Human-machine Agreement. In *Empirical Methods in Natural Language Processing*, pages 1741–1752.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 4(70):213–220.

Richard Craggs and Mary McGee Wood. 2005. Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3):289–295.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Joseph L. Fleiss. 1981. *Statistical methods for rates and proportions*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley.

Laura D. Goodwin and Nancy L. Leech. 2006. Understanding Correlation: Factors That Affect the Size of r. *The Journal of Experimental Education*, 74(3):249–266, April.

Patrick Graham and Rodney Jackson. 1993. The analysis of ordinal agreement data: beyond weighted kappa. *Journal of clinical epidemiology*, 46(9):1055–1062, September.

Kilem Gwet. 2002. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, 2:1–9.

Kilem L. Gwet. 2014. *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Jan Hauke and Tomasz Kossowski. 2011. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30(2):87–93, January.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random er- ror of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis and J. C. Burstein, editors, *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112.

G Frank Lawlis and Elba Lu. 1972. Judgment of counseling process: reliability, agreement, and error. *Psychological bulletin*, 78(1):17–20, July.

Tsun-Jui Liu, Shu-Kai Hsieh, and Laurent PREVOT. 2013. Observing Features of PTT Neologisms: A Corpus-driven Study with N-gram Model. In *Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 250–259.

Zvi Maimon, Adi Raveh, and Gur Mosheiov. 1986. Additional cautionary notes about the Pearson's correlation coefficient. *Quality and Quantity*, 20(4).

Ellis B. Page. 1968. The use of the computer in analyzing student essays. *International Review of Education*, 14(2):210–225, June.

L Perelman. 2013. Critique (ver. 3.4) of mark d. shermis and ben hammer,contrasting state-of-the-art automated scoring of essays: Analysis. *New York Times*.

Donald E. Powers, Jill C. Burstein, Martin Chodorow, Mary E. Fowles, and Karen Kukich. 2002. Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2):103–134.

David M W Powers. 2012. The problem with kappa. In *13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–355.

Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2):3–21.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, 19(3):321–325.

Mark D Shermis and Ben Hamner. 2012. Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual National Council on Measurement in Education Meeting*, pages 14–16.

Advaith Siddharthan and Napoleon Katsos. 2010. Reformulating Discourse Connectives for Non-Expert Readers. In *North American Chapter of the ACL*, pages 1002–1010.

Julius Sim and Chris C Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85(3):257–268, March.

John S Uebersax. 1987. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101(1):140.

Jinhao Wang and Michelle Stallone Brown. 2007. Automated Essay Scoring Versus Human Scoring: A Comparative Study. *The Journal of Technology, Learning and Assessment*, 6(2).

Jinhao Wang and Michelle Stallone Brown. 2008. Automated Essay Scoring Versus Human Scoring: A Correlational Study. *Contemporary Issues in Technology and Teacher Education*, 8(4):310–325.

David M. Williamson, Xiaoming Xi, and Jay F. Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

David M. Williamson. 2009. A Framework for Implementing Automated Scoring. In *Proceedings of the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education*, San Diego, CA.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Xinshu Zhao. 2011. When to use scott's π or krippendorff's α, if ever? In *The annual Conference of the Association for Education in Journalism and Mass Communication*.