A Proposal on Evaluation Measures for RTE

Richard Bergmair
recipient of a DOC-fellowship of the Austrian Academy of Sciences
at the University of Cambridge Computer Laboratory;
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK;
rbergmair@acm.org

Abstract

We outline problems with the interpretation of accuracy in the presence of bias, arguing that the issue is a particularly pressing concern for RTE evaluation. Furthermore, we argue that average precision scores are unsuitable for RTE, and should not be reported. We advocate mutual information as a new evaluation measure that should be reported in addition to accuracy and confidence-weighted score.

1 Introduction

We assume that the reader is familiar with the evaluation methodology employed in the RTE challenge.[1] We address the following three problems we currently see with this methodology.

1. The distribution of three-way gold standard labels is neither balanced nor representative of an application scenario. Yet, systems are rewarded for learning this artificial bias from training data, while there is no indication of whether they could learn a different bias.

2. The notion of confidence ranking is misleading in the context of evaluating a ranking by average precision. The criteria implicitly invoked on rankings by the current evaluation measures can, in fact, contradict those invoked on labellings derived by rank-based thresholding.

3. Language allows for the expression of logical negation, thus imposing a symmetry on the judgements ENTAILED vs. CONTRADICTION. Average precision does not properly reflect this symmetry.

In this paper, we will first summarize relevant aspects of the current methodology, and outline these three problems in greater depth.

The problem of bias is quite general and widely known. Artstein and Poesio (2005) discuss it in the context of Cohen's kappa (Cohen, 1960), which is one way of addressing the problem. Yet, it has not received sufficient attention in the RTE community, which is why we will show how it applies to RTE, in particular, and why it is an especially pressing concern for RTE.

Average precision has been imported into the RTE evaluation methodology from IR, tacitly assuming a great level of analogy between IR and RTE. However, we will argue that the analogy is flawed, and that average precision is not suitable for RTE evaluation.

Then, we will then reframe the problem in information theoretic terms, advocating mutual information as a new evaluation measure. We will show that it addresses all of the issues raised concerning accuracy and average precision and has advantages over Cohen's kappa.

2 The Structure of RTE Data

Let $\mathcal{X}$ be the set of all candidate entailments that can be formed over a natural language of interest, such as English. An RTE dataset $\mathrm{X} \subseteq \mathcal{X}$ is a set of N candidate entailments $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}\}$.

The RTE task is characterized as a classification task. A given candidate entailment x_i can be associated with either a positive class label $\triangle$ (TRUE / YES / ENTAILED) or a negative class label $\bigtriangledown$ (FALSE / NO / NOT ENTAILED), but never both. In the three-way subtask, the positive class, which we will denote as $\boxplus$, is defined as before, but the negative class $\bigtriangledown$ is further subdivided into a class $\boxminus$ (NO / CONTRADICTION) and a class $\Diamond$ (UNKNOWN). To model this subdivision, we define equivalence classes $[\cdot]_3$ and $[\cdot]_2$ on the three-way labels as follows: $[\boxplus]_3 = \boxplus$, $[\Diamond]_3 = \Diamond$, $[\boxminus]_3 = \boxminus$, $[\boxplus]_2 = \triangle$, $[\Diamond]_2 = \bigtriangledown$, and $[\boxminus]_2 = \bigtriangledown$.

The gold standard G for dataset X is then a labelling $\mathrm{G} : \mathrm{X} \mapsto \{\boxplus, \Diamond, \boxminus\}$. We call a candidate

[1] see the reports on RTE-1 (Dagan et al., 2005), RTE-2 (Bar-Haim et al., 2006), RTE-3 (Giampiccolo et al., 2007), the RTE-3 PILOT (Voorhees, 2008), RTE-4 (Giampicolo et al., 2008), and RTE-5 (TAC, 2009)

Proceedings of the 2009 Workshop on Applied Textual Inference, ACL-IJCNLP 2009, pages 10–17,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

entailment x_i a $\triangle$-instance iff $[G(x_i)]_2 = \triangle$, and analogously for the other class labels.

The output $(L, \succ)$ of an RTE system on dataset X also contains such a labelling $L : X \mapsto \{\boxplus, \Diamond, \boxminus\}$, in addition to a strict total order $\succ$ on X representing a ranking of candidate entailments.

2.1 Logical Preliminaries

The notation chosen here is inspired by modal logic. Let's say a candidate entailment x_i were of the logical form $\varphi \to \psi$. The formula "$\Box(\varphi \to \psi)$" would then assert that ψ *necessarily* follows from φ (ENTAILMENT), and the formula "$\Box(\varphi \to \neg\psi)$", which would be equivalent to "$\neg\Diamond(\varphi \wedge \psi)$", would mean that we can *not possibly* have $\varphi \wedge \psi$ (CONTRADICTION). We think of the former as a positive form of necessity ($\boxplus$), and of the latter as a negative form of necessity ($\boxminus$). The formula "$\Diamond(\varphi \to \psi)$" would assert that ψ *possibly* follows from φ (UNKNOWN).

We will have to assume that this negation operator $\neg$ is in fact within the expressive power of the natural language of interest, i.e. "$\varphi \to \neg\psi$" $\in \mathcal{X}$, whenever "$\varphi \to \psi$" $\in \mathcal{X}$. It imposes a symmetry on the two labels $\boxplus$ and $\boxminus$, with $\Diamond$ being neutral.

For example: *"Socrates is a man and every man is mortal; Therefore Socrates is mortal."* This candidate entailment is a $\boxplus$-instance. It corresponds to the following $\boxminus$-instance: *"Socrates is a man and every man is mortal; Therefore Socrates is not mortal"*. But then, consider the $\Diamond$-instance *"Socrates is mortal; Therefore Socrates is a man"*. Here *"Socrates is mortal; Therefore Socrates is not a man"* is still a $\Diamond$-instance.

It is this modal logic interpretation which matches most closely the ideas conveyed by the task definitions (TAC, 2009), and the annotation guidelines (de Marneffe and Manning, 2007). However, for the two-way task, they allude more to probabilistic logic or fuzzy logic, where a candidate entailment is a $\triangle$-instance iff it holds to a higher degree or likelihood or probability than its negation, and a $\triangledown$-instance otherwise.

We believe that either a three-way modal logic entailment task or a two-way probabilistic logic entailment task on its own could make perfect sense. However, they are qualitatively different and not trivially related by equating $\triangle$ with $\boxplus$, and subdividing $\triangledown$ into $\Diamond$ and $\boxminus$.

3 Accuracy & Related Measures

Both the system and the gold standard apply to the dataset X a total labelling L and G respectively, i.e. they are forced to assign their best guess label to every instance. A degree of agreement can be determined as a percentage agreement either on the two-way or the three-way distinction:

$$\mathbb{A}_3(L; G) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left([L(x_i)]_3 = [G(x_i)]_3\right),$$
$$\mathbb{A}_2(L; G) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left([L(x_i)]_2 = [G(x_i)]_2\right),$$

where $\mathbb{1}$ is a counter which takes on a numerical value of one, when the logical expression in its argument is true, and zero otherwise.

The RTE-3 PILOT (Voorhees, 2008) reported some accuracy measures conditioned on gold standard labels as follows:

$$\mathbb{A}'_3(L; G, g) = \frac{\sum_{i=1}^{N} \mathbb{1}\left([L(x_i)]_3 = [G(x_i)]_3 = g\right)}{\sum_{i=1}^{N} \mathbb{1}\left([G(x_i)]_3 = g\right)},$$
$$\mathbb{A}'_2(L; G, g) = \frac{\sum_{i=1}^{N} \mathbb{1}\left([L(x_i)]_2 = [G(x_i)]_2 = g\right)}{\sum_{i=1}^{N} \mathbb{1}\left([G(x_i)]_2 = g\right)}.$$

Assuming the usual analogy with IR, we note that $\mathbb{A}'_2(L; G, \triangle)$ is akin to recall. On the other hand, $\mathbb{A}'_2(G; L, \triangle)$, which conditions accuracy on the system-assigned labels rather than the gold standard labels, is precision.

The conditioned accuracy measures do not provide a single summary statistic as the others do. However, such a summary could be defined by taking the mean across the different labels:

$$\mathbb{A}'_3(L; G) = \frac{1}{3} \sum_{g \in \{\boxplus, \Diamond, \boxminus\}} \mathbb{A}'_3(L; G; g),$$
$$\mathbb{A}'_2(L; G) = \frac{1}{2} \sum_{g \in \{\triangle, \triangledown\}} \mathbb{A}'_2(L; G; g).$$

It is instructive to consider a number of trivial baseline systems. Let $S^{\boxplus}$, $S^{\Diamond}$, and $S^{\boxminus}$, be the systems that uniformly assign to everything the labels $\boxplus$, $\Diamond$, and $\boxminus$, respectively, so that for all i: $L^{\boxplus}(x_i) = \boxplus$, $L^{\Diamond}(x_i) = \Diamond$, and $L^{\boxminus}(x_i) = \boxminus$. Also consider system S^{*}, which assigns labels at random, according to a uniform distribution.

The performance of these systems depends on the distribution of gold-standard labels. The policy at RTE was to sample in such a way that the resulting two-way labels in the gold standard would be balanced. So 50% of all i had $[\mathrm{G}(\mathrm{x}_i)]_2 = \triangle$, while the other 50% had $[\mathrm{G}(\mathrm{x}_i)]_2 = \triangledown$.

This means that all trivial baselines have an accuracy of $\mathbb{A}_2 = \mathbb{A}'_2 = 50\%$. If the data were balanced on the three-way labels, which they are not, we would analogously have $\mathbb{A}_3 = \mathbb{A}'_3 = 33\%$.

When interpreting a two-way accuracy, one would thus expect values between 50% and 100%, where 50% indicates a trivial system and 100% indicates a perfect system. A value of, for example, 70% could be interpreted as-is, mindful of the above range restriction, or the range restriction could be factored into the value by using a linear transformation. One would then say that the accuracy of 70% is 40% of the way into the relevant range of $50\% - 100\%$, and quote the value as a Cohen's Kappa of $\kappa = 0.4$.

3.1 Bias

While the RTE datasets are balanced on two-way gold standard labels, they are not balanced on the three-way gold standard labels. Among the candidate entailments x_i with $[\mathrm{G}(\mathrm{x}_i)]_2 = \triangledown$, in RTE-4, 70% of all x_i had $[\mathrm{G}(\mathrm{x}_i)]_3 = \Diamond$, while only 30% had $[\mathrm{G}(\mathrm{x}_i)]_3 = \boxminus$. In the RTE-3 PILOT, the distribution was even more skewed, at 82%/18%.

So, we observe that $\mathrm{S}^{\boxplus}$ has $\mathbb{A}_3(\mathrm{L}^{\boxplus}; \mathrm{G}) = .500$ and therefore outperforms two thirds of all RTE-3 PILOT participants and one third of all RTE-4 participants. On the other hand, only very few participants performed worse than the random choice system S^{*}, which had $\mathbb{A}_3(\mathrm{L}^{*}; \mathrm{G}) = .394$ on RTE-4. The other trivial systems have $\mathbb{A}_3(\mathrm{L}^{\Diamond}; \mathrm{G}) = .350$, followed by $\mathbb{A}_3(\mathrm{L}^{\boxminus}; \mathrm{G}) = .150$ on RTE-4.

The conditioned accuracies seem to promise a way out, since they provide an artificial balance across the gold standard labels. We have $\mathbb{A}'_3(\mathrm{L}^{\boxplus}; \mathrm{G}) = \mathbb{A}'_3(\mathrm{L}^{\Diamond}; \mathrm{G}) = \mathbb{A}'_3(\mathrm{L}^{\boxminus}; \mathrm{G}) = .33$. But this measure is then counter-intuitive in that the random-choice system S^{*} gets $\mathbb{A}'_3(\mathrm{L}^{*}; \mathrm{G}) = .394$ on RTE-4 and would thus be considered strictly superior to the system $\mathrm{S}^{\boxplus}$, which, if nothing else, at least reproduces the right bias. Another caveat is that this would weigh errors on rare labels more heavily than errors on common labels.

In some form or another the problem of bias applies not only to accuracy itself, but also to related statistics, such as precision, recall, precision/recall curves, and confidence weighted score. It is therefore quite general, and there are three responses which are commonly seen:

1. For purposes of intrinsic evaluation, one can use samples that have been balanced artificially, as it is being done in the two-way RTE task. Yet, it is impossible to balance a dataset both on a two-way and a three-way labelling at the same time.

2. One can use representative samples and argue that the biased accuracies have an extrinsic interpretation. For example, in IR, precision is the probability that a document chosen randomly from the result set will be considered relevant by the user. Yet, for RTE, one cannot provide a representative sample, as the task is an abstraction over a number of different applications, such as information extraction (IE), question answering (QA), and summarization (SUM), all of which give rise to potentially very different distributions of labels.

3. On statistical grounds, one can account for the possibility of random agreement in the presence of bias using Cohen's kappa (Artstein and Poesio, 2005; Di Eugenio and Glass, 2004). We will outline mutual information as an alternative, arguing that it has additional advantages.

4 Average Precision

The purpose of average precision is to evaluate against the gold standard labelling G the system-assigned ranking $\succ$, rather than directly comparing the two labellings G and L.

This is done by deriving from the ranking $\succ$ a series of binary labellings. The i-th labelling in that series is that which labels all instances up to rank i as $\triangle$. A precision value can be computed for each of these labellings, compared to the same gold standard, and then averaged.

More formally, $\succ$ is the strict total ordering on the dataset X which has been produced by the system. Let $\mathrm{x}_j \succeq \mathrm{x}_i$ iff $\mathrm{x}_j \succ \mathrm{x}_i$ or $\mathrm{x}_j = \mathrm{x}_i$. We can then associate with each instance x_i a numeric rank, according to its position in $\succ$:

$$\#_{\succ}(\mathrm{x}_i) = \sum_{j=1}^{\mathrm{N}} \mathbb{1}(\mathrm{x}_j \succeq \mathrm{x}_i).$$

We can then define the cutoff labelling $\succ^{(r)}$ as

$$\succ^{(r)}(\mathrm{x}_i) = \begin{cases} \triangle & \text{if } \#_{\succ}(\mathrm{x}_i) \leqslant r, \\ \triangledown & \text{otherwise;} \end{cases}$$

and average precision as

$$\mathbb{AP}(\mathrm{G}; \succ) = \frac{1}{\mathrm{N}} \sum_{r=1}^{\mathrm{N}} \mathbb{A}'_2\Big(\mathrm{G}; \succ^{(r)}, \triangle\Big).$$

The system-assigned labelling L and the series of ranking-based labellings $\succ^{(r)}$ are initially independent, but, since both accuracy and average precision refer to the same gold standard G, we get the following condition on how L must relate to $\succ$: We call a system output $(\mathrm{L}, \succ)$ sound if there exists a cutoff rank r, such that L equals $\succ^{(r)}$, and self-contradictory otherwise. This is because, for a self-contradictory system output, there does not exist a gold standard for which it would be perfect, in the sense that both accuracy and average precision would simultaneously yield a value of 100%.

So far, we avoided the common terminology referring to $\succ$ as a "confidence ranking", as the notion of confidence would imply that we force the system to give its best guess labels, but also allow it to provide a measure of confidence, in this case by ranking the instances, to serve as a modality for the interpretation of such a best guess.

This is not what is being evaluated by average precision. Here, a system can remain entirely ignorant as to what is a $\triangle$- or a $\bigtriangledown$-instance. System-assigned labels do not enter the definition, and systems are not required to choose a cutoff r to derive a labelling $\succ^{(r)}$. This sort of evaluation is adequate for IR purposes, where the system output is genuinely a ranking, and it is up to the user to set a cutoff on what is relevant to them. As for RTE, it is unclear to us whether this applies.

4.1 Thresholding

In the previous section, we have seen that it is somewhat misleading to see $\succ$ as a confidence-ranking on the labelling L. Here, we argue that, even worse than that, the interpretations of $\succ$ and L may contradict each other. It is impossible for a system to optimize its output $(\mathrm{L}, \succ)$ for accuracy $\mathbb{A}_2\big(\mathrm{G};\mathrm{L}\big)$ and simultaneously for average precision $\mathbb{AP}(\mathrm{G};\succ)$, while maintaining as a side condition that the information state $(\mathrm{L}, \succ)$ remain sound at all times. We show this by indirect argument.

For the sake of contradiction, assume that the system has come up with an internal information state consisting of the ranking $\succ$ and the labelling L, as a best guess. Also assume that this information state is sound.

Let's assume furthermore, again for the sake of contradiction, that the system is now allowed to query an oracle with access to the gold standard in order to revise the internal information state with the goal of improving its performance as measured by accuracy, and simultaneously also improving its performance as measured by average precision.

First, the oracle reveals r, the number of $\triangle$-instances in the gold standard. Let instance $\mathrm{x_i}$ at rank $\#_{\succ}(\mathrm{x_i}) = \mathrm{r}$ be correctly classified, and the instance $\mathrm{x_j}$ at some rank $\#_{\succ}(\mathrm{x_j}) > \mathrm{r}+1$ be incorrectly classified. So we would have $[\mathrm{L}(\mathrm{x_i})]_2 = \mathrm{L}^{(\mathrm{r})}_{\succ}(\mathrm{x_i}) = [\mathrm{G}(\mathrm{x_i})]_2 = \triangle$, and $[\mathrm{L}(\mathrm{x_j})]_2 = \mathrm{L}^{(\mathrm{r})}_{\succ}(\mathrm{x_j}) = \bigtriangledown \neq [\mathrm{G}(\mathrm{x_j})]_2$.

Next, the oracle reveals the fact that $\mathrm{x_j}$ had been misclassified. In response to that new information, the system could change the classification and set $\mathrm{L}(\mathrm{x_j}) \leftarrow \triangle$. This would lead to an increase in accuracy. Average precision would remain unaffected, as it is a function of $\succ$, not L.

However, the information state $(\mathrm{L}, \succ)$ is now self-contradictory. The ranking $\succ$ would have to be adapted as well to reflect the new information. Let's say $\mathrm{x_j}$ were reranked by inserting it at some rank $\mathrm{r}' \leqslant \mathrm{r}$. This would lead to all intervening instances, including $\mathrm{x_i}$, to be ranked down, and thus to an increase in average precision.

But, since $\mathrm{x_i}$ has now fallen below the threshold r, which was, by definition, the correct threshold chosen by the oracle, the system would reclassify it as $[\mathrm{L}(\mathrm{x_j})]_2 = \bigtriangledown$, which now introduces a labelling error. While average precision would not react to this relabelling, accuracy would now drop.

So there are two rather counterintuitive conclusions concerning the simultaneous application of accuracy, average precision, and thresholding. First, accuracy may prefer self-contradictory outputs to sound outputs. Second, when soundness is being forced, average precision may prefer lower accuracy to higher accuracy labellings.

Again, it should be stressed that RTE is the only prominent evaluation scheme we know of that insists on this combination of accuracy and average precision. If we had used precision and average precision, as in IR, the above argument would not hold. Also, in IR, average precision clearly dominates other measures in its importance.

4.2 Logical Symmetry

Besides the above arguments on bias, and on the contradictions between accuracy and average precision under a thresholding interpretation, there is a third problem with the current evaluation methodology. It arises from the symmetry between the classes $\boxplus$ and $\boxminus$ which we introduced in section 2.1. This problem is a direct result of the inherent properties of language and logic, and is, thus, the argument which is most specific to RTE.

Let $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}\}$ be a dataset, and let

$$\neg\mathrm{X} = \{\neg\mathrm{x}_1, \neg\mathrm{x}_2, \ldots, \neg\mathrm{x}_\mathrm{N}\}$$

be the dataset resulting from the application of negation to each of the candidate entailments. Similarly, let $\mathrm{G} : \mathrm{X} \mapsto \{\boxplus, \Diamond, \boxminus\}$ be a gold standard and for all $x \in \mathrm{X}$, let

$$\neg\mathrm{G}(\neg x) = \begin{cases} \boxminus & \text{if } \mathrm{G}(x) = \boxplus, \\ \Diamond & \text{if } \mathrm{G}(x) = \Diamond, \\ \boxplus & \text{if } \mathrm{G}(x) = \boxminus, \end{cases}$$

and analogously for the system-assigned labels L.

Intuitively, we would now expect the following of an evaluation measure: A system that produces the labelling L for dataset X is equivalent, in terms of the evaluation measure, to a system that produces labelling $\neg$L for dataset $\neg$X. This is indeed true for three-way accuracy, where $\mathbb{A}_3\big(\mathrm{G};\mathrm{L}\big) = \mathbb{A}_3\big(\neg\mathrm{G};\neg\mathrm{L}\big)$, but it is not true for two-way accuracy, where the three-way classes are now lumped together in a different way.

Also, this symmetry is not present in average precision, which looks only at positive instances. Since the set of $\triangle$-instances of X and the set of $\triangle$-instances of $\neg$X are disjoint, the two average precisions $\mathbb{AP}(\mathrm{G}; \succ)$ and $\mathbb{AP}(\neg\mathrm{G}; \succ')$, regardless of how $\succ$ relates to $\succ'$, need not be functionally related. – This makes sense in IR, where the set of irrelevant and non-retrieved documents must not enter into the evaluation of a retrieval system. But it makes no sense for the RTE task, where we do need to evaluate systems on the ability to assign a single label to all and only the contradictory candidate entailments.

5 Mutual Information

In this section, we define mutual information as a possible new evaluation measure for RTE. In particular, we return to the problem of bias and show that, like Cohen's kappa, mutual information does not suffer from bias. We will then introduce a new problem, which we shall call degradation. We show that Cohen's kappa suffers from degradation, but mutual information does not. Finally, we will extend the discussion to account for confidence.

Recall that an RTE dataset is a set of N candidate entailments $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}\}$, and let $\mathbf{X}$ be a random variable representing the result of a random draw from this set. Let $\mathbb{P}(\mathbf{X} = \mathrm{x}_i)$ be the probability that x_i comes up in the draw. This could represent, for example, the prior probability that a particular question is asked in a question answering scenario. In the absence of any extrinsically defined interpretations, one could set random variable $\mathbf{X}$ to be uniformly distributed, i.e. $\mathbb{P}(\mathbf{X} = \mathrm{x}_i) = \frac{1}{\mathrm{N}}$ for all i.

This yields a number of further random variables: Let $\mathbf{G}$ and $\mathbf{L}$ be the label $\mathrm{G}(\mathrm{x}_i)$ and $\mathrm{L}(\mathrm{x}_i)$ respectively, assigned to the candidate x_i which has been drawn at random. As usual, we will be interested in their joint distribution, and the resulting marginals and conditionals.

We give the remaining definitions leading to mutual information in Figure 1, and will discuss them by considering the particular contingency table in Figure 2 as an example. It also spells out the information theoretic calculations in detail. Furthermore, we will present corresponding values for Cohen's kappa, which should be easy for the reader to retrace, and thus have been omitted from the Figure for brevity.

The unconditional entropy $\mathbb{H}(\mathbf{G})$ serves as a convenient measure of the hardness of the classification task itself, taking into account the number of labels and their distribution in the gold standard. In the example, this distribution has been chosen to match that of the RTE-4 dataset almost precisely, yielding a value for $\mathbb{H}(\mathbf{G})$ of 1.4277 bits. This indicates that it is much harder to guess the three-way gold standard label of an RTE-4 candidate entailment than it is to guess the two-way label, or the outcome of a toss of a fair coin, which would both have an entropy of exactly 1 bit. On the other hand, due to the skewness of the distribution, it is easier to guess this outcome than it would be if the distribution was uniform, in which case we would have an entropy of 1.5850 bits.

Similarly, we can calculate a conditional entropy $\mathbb{H}(\mathbf{G}|\mathbf{L} = l)$ over a conditional distribution of gold standard labels observed, given that the system has assigned label l to our randomly chosen candidate entailment. In the example, we have calculated a value of 1.0746 bits for $\mathbb{H}(\mathbf{G}|\mathbf{L} = \boxplus)$. So, while the hardness of guessing the correct label without any additional knowledge is 1.4277, it will be easier to guess this label correctly once the system-assigned label is known to be $\boxplus$.

Our best guess would be to always assign label $\boxplus$, which would be successful 50% of the time.

$$
\begin{aligned}
\mathbb{P}(\mathbf{G} = g, \mathbf{L} = l) &= \sum_{i=1}^{\mathrm{N}} \mathbb{P}(\mathbf{X} = \mathrm{x}_i)\ \mathbb{1}\Big(\mathrm{G}(\mathrm{x}_i) = g\ \wedge\ \mathrm{L}(\mathrm{x}_i) = l\Big); && (1)\\
\mathbb{P}(\mathbf{G} = g) &= \sum_{l} \mathbb{P}(\mathbf{G} = g, \mathbf{L} = l) && (2)\\
\mathbb{P}(\mathbf{L} = l) &= \sum_{g} \mathbb{P}(\mathbf{G} = g, \mathbf{L} = l) && (3)\\
\mathbb{P}(\mathbf{G} = g | \mathbf{L} = l) &= \frac{\mathbb{P}(\mathbf{G} = g, \mathbf{L} = l)}{\mathbb{P}(\mathbf{L} = l)}; && (4)\\
\mathbb{H}(\mathbf{G}) &= -\sum_{g} \mathbb{P}(\mathbf{G} = g)\ \log\Big(\mathbb{P}(\mathbf{G} = g)\Big); && (5)\\
\mathbb{H}(\mathbf{G} | \mathbf{L} = l) &= -\sum_{g} \mathbb{P}(\mathbf{G} = g | \mathbf{L} = l)\ \log\Big(\mathbb{P}(\mathbf{G} = g | \mathbf{L} = l)\Big); && (6)\\
\mathbb{H}(\mathbf{G} | \mathbf{L}) &= \sum_{l} \mathbb{P}(\mathbf{L} = l)\ \mathbb{H}(\mathbf{G} | \mathbf{L} = l); && (7)\\
\mathbb{I}(\mathbf{G}; \mathbf{L}) &= \mathbb{H}(\mathbf{G}) - \mathbb{H}(\mathbf{G} | \mathbf{L}). && (8)
\end{aligned}
$$

Figure 1: definitions for mutual information $\mathbb{I}(\mathbf{G}; \mathbf{L})$

20 (45)	25 (0)	5	$\mathbb{P}(\mathrm{G} = \boxplus)$ $= .5$
9 (27)	18 (0)	9	$\mathbb{P}(\mathrm{G} = \diamondsuit)$ $= .36$
1 (8)	7 (0)	6	$\mathbb{P}(\mathrm{G} = \boxminus)$ $= .14$
$\mathbb{P}(\mathrm{L} = \boxplus)$ $= .3$ (.8)	$\mathbb{P}(\mathrm{L} = \diamondsuit)$ $= .5$ (0)	$\mathbb{P}(\mathrm{L} = \boxminus)$ $= .2$ (.2)	$\mathrm{N} = 100$

$$
\begin{aligned}
-\mathbb{H}(\mathbf{G}) &= .5\ \log_2(.5)\\
&\quad + .36\ \log_2(.36)\\
&\quad + .14\ \log_2(.14)\\
&\quad = -1.4277
\end{aligned}
$$

$$
\begin{aligned}
-\mathbb{H}(\mathbf{G} | \mathbf{L} = \boxplus) &= \frac{20}{30}\ \log_2(\frac{20}{30})\\
&+ \frac{9}{30}\ \log_2(\frac{9}{30})\\
&+ \frac{1}{30}\ \log_2(\frac{1}{30})\\
&= -1.0746
\end{aligned}
\qquad
\begin{aligned}
-\mathbb{H}(\mathbf{G} | \mathbf{L} = \diamondsuit) &= \frac{25}{50}\ \log_2(\frac{25}{50})\\
&+ \frac{18}{50}\ \log_2(\frac{18}{50})\\
&+ \frac{7}{50}\ \log_2(\frac{7}{50})\\
&= -1.4277
\end{aligned}
\qquad
\begin{aligned}
-\mathbb{H}(\mathbf{G} | \mathbf{L} = \boxminus) &= \frac{5}{20}\ \log_2(\frac{5}{20})\\
&+ \frac{9}{20}\ \log_2(\frac{9}{20})\\
&+ \frac{6}{20}\ \log_2(\frac{6}{20})\\
&= -1.5395
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{H}(\mathbf{G} | \mathbf{L}) &= .3 * 1.0746\\
&+ .5 * 1.4277\\
&+ .2 * 1.5395\\
&= 1.3441
\end{aligned}
\qquad
\begin{aligned}
-\mathbb{H}(\mathbf{G} | \mathbf{L}' = \boxplus) &= \frac{45}{80}\ \log_2(\frac{45}{80})\\
&+ \frac{27}{80}\ \log_2(\frac{27}{80})\\
&+ \frac{8}{80}\ \log_2(\frac{8}{80})\\
&= -1.3280
\end{aligned}
\qquad
\begin{aligned}
\mathbb{H}(\mathbf{G} | \mathbf{L}') &= .8 * 1.3280\\
&+ .2 * 1.5395\\
&= 1.3703
\end{aligned}
$$

Figure 2: example contingency table and entropy calculations

But, among the cases where the system in Figure 2 has assigned label $\boxplus$, this would be an even better guess. It would now be correct 66% of the time. We have gained information about the gold standard by looking at the system-assigned label.

5.1 Bias

The conditional entropy $\mathbb{H}(\mathbf{G}|\mathbf{L})$ is the expected value of the conditional entropy $\mathbb{H}(\mathbf{G}|\mathbf{L} = l)$ across all possible labels l, when, as before, we draw a candidate entailment at random.

One very noteworthy property of this measure is that all of the baseline systems we considered, i.e. systems assigning constant labels, or systems assigning labels at random, would have $\mathbb{H}(\mathbf{G}|\mathbf{L}) = \mathbb{H}(\mathbf{G})$, since the distribution of gold standard labels given the system labels, in all of these cases, is the same as the prior distribution. Furthermore, $\mathbb{H}(\mathbf{G}) = 1.4277$ is, in fact, an upper bound on $\mathbb{H}(\mathbf{G}|\mathbf{L})$. All the trivial baseline systems would perform at this upper bound level.

At the other extreme end of the spectrum, consider a perfect contingency table, where all the non-diagonal cells are zero. In this case all the conditional entropies $\mathbb{H}(\mathbf{G}|\mathbf{L} = l)$ would be entropies over delta distributions concentrating all probability mass on a single label. This would yield a value of $\mathbb{H}(\mathbf{G}|\mathbf{L}) = 0$, which is a lower bound for any entropy. – For Cohen's kappa we would have $\kappa = 1$.

The system producing our contingency table performs worse than this ideal but better than the baselines, at $\mathbb{H}(\mathbf{G}|\mathbf{L}) = 1.3441$. One can subtract $\mathbb{H}(\mathbf{G}|\mathbf{L})$ from the upper bound $\mathbb{H}(\mathbf{G})$ to obtain the mutual information $\mathbb{I}(\mathbf{G};\mathbf{L})$. It is the information gained about $\mathbf{G}$ once the value of $\mathbf{L}$ is revealed. It is obviously still bounded between 0 and $\mathbb{H}(\mathbf{G})$, but is somewhat more intuitive as an evaluation measure, as it restores the basic intuition that larger values indicate higher performance. – Due to a surprising result of information theory it also turns out that $\mathbb{I}(\mathbf{G};\mathbf{L}) = \mathbb{I}(\mathbf{L};\mathbf{G})$. This symmetry is another property one would intuitively expect when comparing two labellings G and L to each other, and is also present for accuracy and kappa.

We can compare the behaviour of this measure to that of accuracy. The accuracy of our example system is simply the sum of the diagonal contingency counts, so it scores at 44%, compared to 50% for the baseline that always assigns label $\boxplus$. The new bias-aware framework provides a quite different point of view. We would now note that the example system does provide $\mathbb{I}(\mathbf{L};\mathbf{G}) = 0.0836$ bits worth of information about $\mathbf{G}$, showing an agreement of $\kappa = 0.1277$, compared to zero information and $\kappa = 0$ agreement for the baseline.

5.2 Degradation

The numbers in the example have been chosen so as to illustrate a problem we call degradation. The conditional distribution $\mathbb{P}(\mathbf{G} = g|\mathbf{L} = \Diamond)$ is the same as the unconditional distribution $\mathbb{P}(\mathbf{G} = g)$, so when it turns out that $\mathbf{L} = \Diamond$, no additional information has been revealed about $\mathbf{G}$. But in information theoretic terms, it is considered good to know when exactly we know nothing.

What happens if we conflate the labels $\Diamond$ and $\boxplus$ in the system output? In Figure, 2, the numbers in brackets illustrate this. Previously, the system assigned label $\boxplus$ in 30% of all cases. In those cases, the system's choice was relatively well-informed, as $\boxplus$ actually turned out to be the correct gold standard label 66% of the time. But now, with the labels conflated, the system chooses $\boxplus$ in 80% of the cases; a choice which is now much less well-informed, as it is correct only 45% of the time.

Mutual information shows a drop from 0.0836 bits down to 0.0262. On the other hand, accuracy increases from 44% to 51%, and Cohen's kappa also increases from 0.1277 to 0.1433. But this is clearly counter-intuitive. Surely, it must be a bad thing to conflate a well-informed label with a less well-informed label, thus obscuring the output to less certainty and more guesswork.

5.3 Confidence Ranking

One final issue that has still remained unaddressed is that of confidence ranking. This takes us back to the very first probabilistic notion we introduced, that of a probability distribution $\mathbb{P}(\mathbf{X} = \mathrm{x}_i)$ governing the choice of the test-instances x_i. The uniform distribution we suggested earlier results in all instances carrying equal weight in the evaluation.

But for some applications, it makes sense to give the system some control over which test-instances it wants to be tested on, independently of the question of what results it produces for that test. – So, from a probabilistic point of view, the most natural take on confidence would be to have the system itself output the values $\mathbb{P}(\mathbf{X} = \mathrm{x}_i)$ as confidence weights.

This would affect $\mathbb{H}(\mathbf{G})$, which we previously introduced as a measure of the difficulty of the task faced by the system. But now, the system has some control over what task it wants to try and solve. In an extreme scenario, it could concentrate all its confidence mass in a single instance. Another system might force itself to give equal weight to every instance. Clearly, these are two very different scenarios, so it seems natural that, as soon as the issue of confidence enters the scene, the evaluation has to consider two dimensions. The unconditional entropy $\mathbb{H}(\mathbf{G})$ would have to be reported for every system, together with the mutual information $\mathbb{I}(\mathbf{L};\mathbf{G})$. While $\mathbb{H}(\mathbf{G})$ would measure how effective a system was at using its confidence weighting as a tool to make the task easier on itself, $\mathbb{I}(\mathbf{L};\mathbf{G})$ would measure how successful the system ultimately was at the task it set for itself.

The example of a system concentrating all of its confidence mass in a single instance shows that the ability to freely choose $\mathbb{P}(\mathbf{X} = \mathrm{x}_i)$ might not fit with realistic application scenarios. This leads to the idea of confidence ranking, where a system could only rank, not weigh, its decisions, and it would be up to the evaluation framework to then assign weights according to the ranks.

For example, one could let

$$\mathbb{P}(\mathbf{X} = \mathrm{x}_i) = \frac{\mathrm{N} + 1 - \#_{>}(\mathrm{x}_i)}{(\mathrm{N}+1) * (\mathrm{N}/2)}.$$

This would assign a weight of N to the highest-ranked instance, a weight of N − 1 to the next, and continue in this manner down to the instance at rank N, which would get weight 1. The denominator in the above expression then serves to normalize this weighting to a probability distribution. Note that, in principle, nothing speaks against using any other series of weights. Perhaps further investigation into the application scenarios of RTE systems will provide an extrinsically motivated choice for such a confidence weighting.

6 Final Recommendations

Ultimately, our proposal boils down to four points, which we believe are well-supported by the evidence presented throughout this paper:

1. Additional clarification is needed as to the logical definitions of the two-way and the three-way distinction of entailment classes.

2. Accuracy and related evaluation measures suffer from bias, and thus scores of theoretical baselines must be reported and compared to system scores. These include random choice and choice of a constant label.

3. Average precision scores are misleading and should not be reported. The confidence-weighted score that has been dropped after RTE-1 would be preferable to average precision, but still suffers from bias.

4. Mutual information should be reported, in addition to accuracy and possibly confidence-weighted score, to account for bias and the degradation problem.

Acknowledgments

I would like to thank the anonymous reviewers and my colleague Ekaterina Shutova for providing many helpful comments and my supervisor Ann Copestake for reading multiple drafts of this paper and providing a great number of suggestions within a very short timeframe. All errors and omissions are, of course, entirely my own. I gratefully acknowledge financial support by the Austrian Academy of Sciences.

References

Ron Artstein and Massimo Poesio. 2005. Kappa3 = alpha (or beta). Technical Report CSM-437, University of Essex Department of Computer Science.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment (RTE-2).*

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

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In Ido Dagan, Oren Glickman, and Bernardo Magnini, editors, *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (RTE-1).*

Marie-Catherine de Marneffe and Christopher Manning. 2007. Contradiction annotation. http:// nlp.stanford.edu/ RTE3-pilot/ contradictions.pdf.

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

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognising textual entailment challenge. In *Proceedings of the Workshop on Textual Entailment and paraphrasing (RTE-3).*

Danilo Giampicolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *Preproceedings of the Text Analysis Conference (TAC).*

TAC. 2009. Tac2009 rte-5 main task guidelines. http:// www.nist.gov/ tac/ 2009/ RTE/ RTE5_Main_Guidelines.pdf.

Ellen M. Voorhees. 2008. Contradictions and justifications: Extensions to the textual entailment task. In *Proceedings of ACL-08: HLT*, pages 63–71.