Inter-rater Agreement Measures and the Refinement of Metrics in the PLATO MT Evaluation Paradigm

Keith J. MILLER
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102
keith@mitre.org

Michelle VANNI
U.S. Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20785
mvanni@arl.army.mil

Abstract

The PLATO machine translation (MT) evaluation (MTE) research program has as a goal the systematic development of a predictive relationship between discrete, well-defined MTE metrics and the specific information processing tasks that can be reliably performed with MT output. Traditional measures of quality, informed by International Standards for Language Engineering (ISLE), namely, clarity, coherence, morphology, syntax, general and domain-specific lexical robustness, and named-entity translation, as well as a DARPA-inspired measure of adequacy are at the core of the program. For robust validation, indispensable for refinement of test and guidelines, we conduct tests of inter-rater reliability on the assessments. Here we discuss development and report on results of our inter-rater reliability tests, focusing on results for Clarity and the Coherence, the first two assessments in the PLATO suite, and we discuss our method for iteratively refining our linguistic metrics and the guidelines for applying them within the PLATO evaluation paradigm. Finally, we discuss reasons why kappa might not be the best measure of inter-rater agreement for our purposes, and suggest directions for future investigation.

1 Introduction

In this paper, we report on achievement of validation of two MT output quality tests within the MTE research program called Predictive Linguistic Assessments of Machine Translation Output (PLATO). In earlier work within this program, we reported on preliminary validation testing on English output of MT systems for the structurally dissimilar input languages of Spanish and Japanese (Miller & Vanni, 2001; Vanni & Miller, 2002; Miller 2004). Our overall research plan entails investigation of correlations between score clustering patterns and tasks for which the MT output found in the cluster has been deemed suitable.

This phased research approach includes the selection of assessments from the Framework for Evaluation of MT in ISLE (FEMTI), design validation and, association of patterns of scores with information processing tasks performable on the output. Since the intent is to automate the scoring system, this work can also be viewed as preliminary phases of algorithm design for automated scoring of MT output.

The PLATO test suite includes assessments of clarity and coherence as well as measures of syntax, morphology, and lexical coverage. Clarity is measured on a scale from unintelligible to meaningful. The coherence metric draws on Mann and Thompson's RST (1981), and is based on impressions of the overall dynamic of a discourse. Scores for syntax are based on the minimal number of corrections needed to render a sentence grammatical; likewise, the morphology scores are based on the rate of word formation errors present in the output text.

We hypothesize that there exists a predictive relationship between these discrete, well-defined linguistic features (as measured by a set of scores for theses quality metrics) and specific information processing tasks that can be reliably performed with the output of a given MT system. We characterize MT output quality in functional terms while responding to the established desiderata for MTE, that it be reliable, replicable, and automatic. Thus, the intended outcomes are (1) a system for classifying MT output in terms of the information processing functions it can serve and (2) indicators for research and development directions to aid developers of MT systems. These indicators will assist them in developing MT that produces output of a quality suitable to serve a specific information processing function.

The inter-rater agreement measures on Clarity and Coherence metrics provide a basis for correlation with independently-derived measures of usefulness for downstream information processing tasks such as those outlined in Doyon, Taylor, & White (1999).

After a brief review of the current state of automated MT Evaluation, we review the tests and the testing process. We then show how different

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Inter-rater Agreement Measures and the Refinement of Metrics in the PLATO MT Evaluation Paradigm	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **MITRE Corporation,7515 Colshire Drive,McLean,VA,22102-7539**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

types of inter-rater agreement measures can reflect the assessors' level of consensus on a given output. In general, the greater the level of consensus, the greater the validity. Our context views validity as a function of principally (1) the consistency with which the test criteria can be applied and (2) the ease with which tests can be applied to varying problematic output, and also considers (3) the extent to which the tests might be automated in later stages of the work.

The multidimensionality of the PLATO metric suite permits the identification of specific error types. Output supporting a given task is subject to cross-feature correlations which include input complexity levels, output characteristics (scores on PLATO linguistic assessments), and automatic metrics.

2 Background

Viewed in its entirety, our research is situated at the juncture of automated MT evaluation and task-based MT evaluation. It also crucially incorporates notions of linguistic quality that are not necessarily intrinsically present in either of these two approaches.

2.1 Automated MT Evaluation

Automated MTE has become a popular area of research since the advent of the n-gram-based BLEU metric (Papineni, et al., 2001). Due to concerns about the semantics of the output which BLEU was judging as "good," variations using, for example, recall (Lavie, et al. 2004), precision and recall (Melamed et al., 2003), saliency measures (Babych and Hartley, 2004), and weighted bigrams (Lin and Och, 2004) have now been developed with authors reporting ever higher correlations with human judgments.

Attention to the linguistic features of output, however, was first given by Jones and Rusk (2000) who compared, using the K-Nearest Neighbor (KNN) algorithm, a set of linguistic test scores for output to a set of the same tests' scores for naturally-occurring target-language text. However, scores on the *ad hoc* tests were compared with scores for human-produced text, possibly differing in type or domain.

Automatic metrics are not designed, however, to provide direction to R&D. As a step towards addressing this diagnostic gap, Papineni, *et al.* (2002) coupled BLEU with the NEE named-entity evaluation tool. Results on DARPA 1994 MTE data revealed correlations with human judgments of fluency and adequacy in the .85 to .94 range, and higher. Note that the translation of named entities is directly related to the use of MT output for information extraction and retrieval purposes. This makes the NEE result an important first contribution in the direction of automatic scoring that provides a comprehensive picture not only of the performance of the system but also of the tasks performable on the output.

2.2 Task-based MT Evaluation

Church and Hovy (1993) proposed that MTE take an approach that gives credit to an MT system for how it serves information processing. This direction has informed activities both in the Expert Advisory Group on Language Engineering Standards (EAGLES) and the ISLE proposals for MTE.

Task-based evaluation evolved from the tradition of black-box evaluation. One of the most widely-cited large scale applications of this approach is the DARPA methodology (White and O'Connell 1994) which measured fluency and accuracy on a 5-point scale. Using scores from the DARPA evaluation series and a set of translation-dependent information processing tasks, experiments were performed to rank tasks on a scale from more to less tolerant of MT output errors (White and Taylor 1998; Taylor and White 1998; Doyon, Talbot and White 1999). More recently, a human-based question-answering task has been explored on outputs of input texts of varying complexity (Jones et al, 2005). Association of task-based evaluation of the type described above with automated metrics has begun to be investigated as well (Weinberg, 2005).

Our goal is to be able to make determinations of which tasks a human analyst or automatic process can perform on output with specific linguistic properties. In selecting specific ISLE features, we recognize that language-dependent tasks vary in their tolerance of error and hypothesize that clustering patterns among the sets of scores will reflect variations along these usability dimensions. In order for clustering based on these linguistically-based metrics to be effective, we must first ensure the quality of the metrics themselves. It is with this goal that we undertook the phased approach to metric definition and refinement described in the remainder of this paper.

3 Data and Methods

3.1 Data

The system output used for these experiments was DARPA 4Q94 Spanish and Japanese MTE data[1]. For each language, three input texts were

[1] The main objective of this part of our research is MT *evaluation.* As such, it is not necessarily focused on differentiating only between the most current versions of operational systems. Thus, as in our previous work in this program, we make use of the valuable resource of the DARPA MT evaluation output, for a portion of which task usability data is available. We are, however, conserving the crucial resource of task-tagged MT output, which is the result of a large human-intensive effort, for a later step in our research, first validating metrics on other data from this corpus before applying them to the actual task-based data set.

selected. Each input had been run through three different MT systems for that language pair, either Spanish-English or Japanese-English. So as not to create bias in favor of a system which may perform particularly well on a given type of input, we selected only two outputs, each from a different system, for each input. In this way, each system and each input was given two chances per phase to be assessed (see Table 1).

Lg - Doc	MT Sys	Output	Phase 1&2
Sp-01	MT01	sp01mt01	*
	MT02	sp01mt02	
	MT03	sp01mt03	*
Sp-02	MT01	sp02mt01	
	MT02	sp02mt02	*
	MT03	sp02mt03	*
Sp-03	MT01	sp03mt01	*
	MT02	sp03mt02	*
	MT03	sp03mt03	
Ja-01	MT01	ja01mt01	*
	MT02	ja01mt02	
	MT03	ja01mt03	*
Ja-02	MT01	ja02mt01	
	MT02	ja02mt02	*
	MT03	ja02mt03	*
Ja-03	MT01	ja03mt01	*
	MT02	ja03mt02	*
	MT03	ja03mt03	

Table 1: MT Output Data for Assessments

Twenty assessors were identified from a pool of candidates comprised of linguistics students, professional copy editors, and teachers of language, to include English as a Second Language, to perform the PLATO Assessments. In Phase One, each assessor saw a set of outputs consisting of six texts randomly chosen from the twelve available MT outputs. In Phase Two, the assessors saw the complementary set of texts. The order of viewing was randomized separately for each phase.

3.2 Methods

An essential element of this experiment was the establishment of guidelines which were iteratively refined throughout the steps of familiarization, presentation of examples, working through of examples, and practice testing.

The features from the ISLE framework that we chose to include in our scoring suite are the following: clarity, coherence, syntax, morphology, and dictionary update/terminology. Criteria for selection of the ISLE features to be tested on the output (the seven assessments) was described in earlier work (Miller & Vanni, 2001). Inter-rater agreement measurements for the Clarity and Coherence assessments and the role of inter-rater agreement measures in the iterative refinement of the metrics are reported on here.

3.2.1 Assessing Clarity

Our Clarity assessment metric for PLATO Phase One ranges between 0 and 3, and can be summarized as in Table 2, below.

Score	Criteria
0	meaning of sentence is not decipherable, even after some reflection
1	meaning of part of the sentence is clear after some reflection
2	meaning of entire sentence is clear after some reflection
3	meaning of entire sentence is perfectly clear on first reading

Table 2: PLATO Phase 1 Clarity Measures

Since the feature of interest for this particular metric is clarity and not fidelity, it is sufficient that some clear meaning is expressed by the sentence and not that that meaning reflect the meaning of the input text. Thus, no reference to the source text or reference translation is permitted. Likewise, for this measure, the sentence need neither make sense in the context of the rest of the text nor be grammatically well-formed, since these features of the text are measured by the Coherence and Syntax tests, respectively. Thus, the clarity score for a sentence is basically a snap judgment of the degree to which some discernible meaning is conveyed by that sentence.

3.2.2 Assessing Coherence

Because coherence is a high-level feature that operates at the supra-sentential level, the goal of this metric is to reflect a general impression of the overall dynamic of the discourse. So, while the coherence of the output texts is assessed using a measure that draws on Mann and Thompson's (1981) Rhetorical Structure Theory (RST), we do not assign RST functions to sub-sentential units; rather, as in our previous work with this assessment we chose the sentence as the unit of evaluation, and scored this feature as the percentage of sentences to which some RST function could be assigned in relation to the discourse as a whole.

We thus apply RST in a very general manner. For our purposes, just as for meaning in the Clarity

test, it matters only whether some logical function be determined for each sentence, not necessarily the "correct" one. RST definitions are used simply to constrain the set of functions that can possibly be assigned to an output sentence. However, function definitions overlap and, in practice, some of the distinctions are too fine-grained for the coarse MT output. Assessors were instructed that if *any* RST function seemed appropriate for the sentence being evaluated, that sentence should receive a positive coherence score. The overall coherence score for an output text, then, is the proportion of sentences to which an RST function can be assigned.

4 Inter-Rater Reliability Measures

We now turn to an examination of the results of calculating interrater reliability measures on the first two of the PLATO metrics – that is the metrics for clarity and coherence. First we consider two measures of agreement. Then, since our principal reason for calculating inter-rater reliability was to determine the reliability of the metric, and to be able to measure increases in the reliability as the metric is refined in subsequent phases of testing, we next discuss potential ways of increasing the kappa value for these metrics. In that discussion, we focus particularly on actions we have taken to increase inter-rater reliability for these metrics in subsequent phases of experimentation and operational evaluation. Finally, we consider reasons that kappa might not be the best measure of inter-rater reliability for evaluation of these machine translation evaluation metrics, and suggest some possible alternative measures of agreement.

4.1 Joint Probability of Agreement

It is possible to compute the joint agreement between any two assessors by calculating the proportion of times that their ratings coincide. For purposes of the clarity test, then, we could simply count the number of times Assessor A assigns a 0 that Assessor B also assigns a 0, and then do the same for the other ratings: {1, 2, 3}. This sum, divided by the total number of ratings results in the joint probability of agreement between Assessor A and Assessor B. This provides an idea of the *absolute* agreement between the two assessors, and would be appropriate if the rating values being assigned were completely categorical in nature. However, recall that the Clarity ratings in PLATO are actually on a graded *scale* from 0 to 3, and as such, disagreement between assessors involving a rating of 0 and a rating of 1 for the same sentence is not the same as a disagreement between assessors involving a rating of 0 and a rating of 3. To account for this feature of the ratings, it is possible to consider degrees of agreement, rather than treating agreement as a binary, all-or-nothing proposition. The weighted version of this calculation allows the assignment of partial credit for items for which there is not an absolute agreement. While it is possible to use a Jaccard coefficient or a Dice coefficient as a weighting function when an assessor is allowed to select a set of labels to apply to each observation, in the case of clarity, which is an ordered categorical variable, we simply weight agreement by the ratio of the absolute difference to the range. This weighting, which also can be applied to the coherence metric[2], a dichotomous variable, can be seen in Figure 1, in which the variables *i* and *j* cover the range of values for the assessment in question.

$$\sum_{i=\min}^{\max}\sum_{j=\min}^{\max}\omega_{i,j}p_{i,j}$$

$$\omega_{i,j}=\frac{|i-j|}{\max-\min}$$

$p_{i,j}$ ≡ proportion of observations in the cell at row i, column j

$p_{i,\bullet}$ ≡ proportion of observations in row i

$p_{\bullet,j}$ ≡ proportion of observations in column j

Figure 1: Weighted Joint Probability of Agreement

4.1.1 Joint Probability of Agreement for Clarity

The calculation outlined above for joint agreement between two raters can be extended to multiple raters by computing the mean of the pairwise joint probabilities. For this calculation, we excluded assessors practice texts, which comprised practice on both human and machine translation outputs designed to train the assessors in application of the metric and thus maximize agreement. Also excluded from this analysis were items for which an assessor did not provide a response, because it was not possible to determine whether this lack of response was intentional or accidental, and further, there was no way to meaningfully include these items in the measurement of agreement. The box plot showing the resultant mean value for joint agreement on the clarity metric is shown in Figure 2 below. Note that despite a couple of outliers at ~ 0.59, and a

[2] Although weighting *can* be applied to the coherence metric, since this is a binary measure, this particular weighting calculation is vacuous. This is because any difference in ratings between raters would be a difference at the two extremes of the range of allowed values, and thus be assigned a weight of 0, which results in a calculation identical to that for absolute agreement.

range with a lower bound at about 0.63, the upper range extends to 0.90, with the mean joint agreement at approximately 0.8, a respectable value.

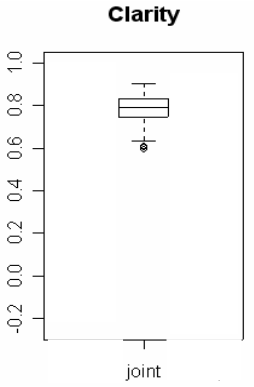

Figure 2: Joint Agreement for PLATO Clarity Metric

4.1.2 Kappa Coefficient for Clarity

It has become a *de* facto standard in the computational linguistics community to employ Cohen's kappa coefficient to quantify agreement between two raters. As is the case for the joint agreement measure in the previous section, extension of kappa to multiple raters is typically done by computing the mean of the pairwise kappa coefficients. The range of the kappa coefficient is from -1 to +1.

Also analogously to the joint probability of agreement, it is possible to compute a weighted version of kappa (Cohen, 1968). The computation for weighted kappa is as follows

$$\text{weighted kappa} = \frac{\text{joint\_sum} - \text{indep\_sum}}{1 - \text{indep\_sum}}$$

Figure 3: Weighted Kappa Computation

Note that kappa takes into account the joint probability of agreement, as defined in Figure 1, and adjusts it based on the independent probability of agreement, that is the probability that agreement is due to chance factors. Using the same calculation for weighting as in Figure 1 above, the independent probabilities are computed as shown in Figure 4.

$$\text{indep\_sum} = \sum_{i=\min}^{\max} \sum_{j=\min}^{\max} \omega_{i,j} p_{i,\bullet} p_{\bullet,j}$$

Figure 4: Computation of Independent Probabilities for Weighted Kappa

The relationship between the joint probability of agreement, the independent probability, and the resulting kappa coefficient is illustrated in Figure 5. Note that although the joint agreement was considerable (0.80), the independent probability was also quite high, resulting in a low kappa value of 0.37.

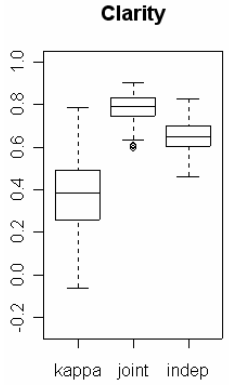

Figure 5: Kappa Coefficient for PLATO Clarity Metric

4.1.3 Joint Agreement and Kappa Coefficient for Coherence

There was a similar situation in the kappa calculation for the coherence metric. Although the mean joint agreement was not quite as remarkable as that for the clarity metric at 0.66, the independent probability was still correspondingly high (0.54), resulting in a low kappa value (0.25). This information is indicated in Figure 6, which also shows that, the low kappa value notwithstanding, some assessors did attain perfect agreement. This reinforces the heavy influence of the independent probabilities on the computation of kappa.

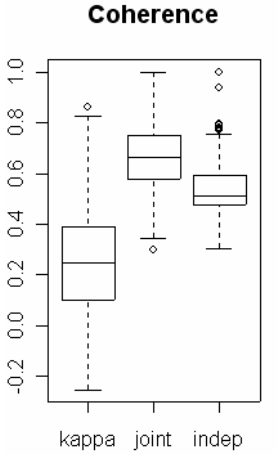

Figure 6: Kappa Coefficient for PLATO Coherence Metric

5 Refinement of Metrics

Although the overall joint agreement on the clarity metric was quite good, and that on the clarity metric was nearing acceptability, the values

for the more stringent measure of inter-rater reliability, kappa, were not as high as we had hoped. Given this situation, we examined ways in which we could increase the reliability of the metrics, and thus correspondingly increase the value of the kappa coefficient in subsequent phases of experimentation. We have previously noted that the subtraction of the products of independent probabilities makes the kappa statistic very sensitive to the marginal distribution of ratings.

As an extreme example, if one assessor always assigns the same rating, there is no way to score above zero. Conversely, the more evenly the assessors' values are distributed across the categories, the better, and the less influence on kappa. In reality, and as often happens, if there is a particular category that gets a disproportionate number of the values (or conversely many fewer values than the other categories), this will have a heavy influence on the value of the resultant kappa coefficient.

Looking again, then, at the computation for weighted kappa in Figure 3, we can see that it is possible to increase kappa by either (1) increasing the joint probability, or (2) decreasing the independent probability.

5.1 Increasing the Joint Probability through the use of Entropy Calculations

Low joint agreement between assessors would simply indicate that assessors are often assigning differing ratings for the same sentence. This would likely be due to poor assessment guidelines or training. In the case of the Clarity and Coherence metrics, the joint agreement was not globally low, yet, even for Clarity, it was still possible to achieve some gains. It was thus necessary to determine which assessment items were most contributing to lowering the mean joint agreement. This was done by computing the entropy of the ratings assigned to each item, and to concentrate on those items with the highest entropy, or, those that had been assigned the largest mix of ratings. The assessment designers examined these high-entropy items, with the aim of discovering any flaws or ambiguities in the assessment guidelines that might have led to the wide range of ratings for the items. The goal of this activity was to recalibrate the assessment guidelines for these metrics for future phases of PLATO experimentation.

For each high-entropy item identified, the assessment designers accessed the assessment item through the assessment interface, and independently rated the item according to the current guidelines. They then discussed any discrepancies among their ratings, and determined which were genuine disagreements as to the quality of the text with respect to the linguistic characteristic being measured, and which were due to vagueness in the assessment guidelines. For the latter, they discussed how the guidelines might be refined in order to ensure greater certainty in rating. Furthermore, they were able to access the assessors' comments, both at the sentence level and at the text level, for each assessment metric being refined. These comments were invaluable in the discussion of possible refinements of the metrics.

[[Note to reviewers: the final version of the paper, if accepted, will contain examples of MT output containing 'high-entropy' and 'low-entropy' items (items with entropy = 0, indicating total agreement among all assessors), as well as examples of comments from assessors that were useful in revising evaluation guidelines.]]

This procedure led to changes in the guidelines for the evaluation metrics that should increase the joint probability of agreement, and thus the kappa coefficient, in future phases of PLATO work.

5.2 Increasing the Joint Probability by the Lowering of Independent Probabilities

We note above that the independent probabilities play a large role in the computation of the kappa statistic when there are categories (i.e. ratings) that account for a disproportionate amount of the data, whether they be ratings that are almost always used on one hand, or ratings that are seldom to never used on the other hand. Thus, in order to minimize the effect of the independent probabilities on kappa, we wish to have a rating scheme across which values are evenly distributed. In practice, there are at least two reasons why this might not be the case.

First, it may be that the object of evaluation is of a uniform quality. That is, if all of the MT output being evaluated is completely clear, we would expect nearly every sentence to be assigned a Clarity score of 3, and almost no sentence to receive a score of 0 or 1. This would lead to a high contribution of the probability of chance agreement. In the field of MT evaluation, this should not be the case. The goal of evaluation being to distinguish between capabilities of different systems, we would not expect all systems to perform equally well on all metrics. In the event that a metric does produce uniformly high or low values across systems, that metric is not useful for discriminating between systems of differing capabilities at an appropriate level of granularity, and should be rethought.

A second cause for a particular category accounting for a disproportionate amount of data

may be that the rating values themselves are not uniformly distributed, or that the rating guidelines give preference (or dispreference) to one particular rating over others. If this is the case, ratings will be non-uniformly distributed, again leading to high contribution of the independent probabilities in the kappa computation.

In discussing the revised guidelines for Phase Two of the PLATO Clarity assessment with the PLATO assessors, and during the working of examples to test the new guidelines, it became apparent that there was a need to further refine the clarity rating scale itself. It was the opinion of the assessors that, on the 4-point clarity rating scale, the distance between ratings 1 and 2 was greater than the distance between 0 and 1 or the distance between 3 and 4. Thus, after some follow-on discussion and further guideline refinement, the scale was expanded to a 5-point scale, in hopes that in future assessments, this would more evenly distribute the values across the scale, thus strengthening the assessment, resulting in a lower independent probability of agreement, and a higher kappa value[3].

Similar procedures were followed to refine the metrics for the remaining PLATO assessments as well.

5.3 An Outstanding Possibility

In considering ways to increase kappa by examining the equation in Figure 3, we made one assumption that caused us to fail to take one possibility into account. It is possible that kappa might not be the best or only way to measure agreement among assessors. We may, for example, benefit from an agreement statistic that is not affected so heavily by the marginal distribution of ratings.

This applies to the PLATO evaluation suite, which contains metrics which have between two and five categories per assessment for most assessments. As a case in point, in the coherence assessment, PLATO makes a binary distinction, and many more sentences are generally judged to be coherent (1) than non-coherent (0), depending of course on the system being evaluated. Hence, the marginal probabilities play a large role. We may ask if there is a better statistic than kappa to use in this case.

[3] Note that this will only be the case if the independent probability is lowered without the expense of lowering the joint probability of agreement. These factors were weighed in making the decision to expand Clarity to a five-point scale.

6 Conclusions and Future Work

Specifically for the Clarity test, and for other PLATO metrics as well, we plan to investigate latent class modelling as an alternative to kappa for measuring inter-rater agreement (Agresti, 2002; Uebersax, 2003). We have also noted that it is possible to compute the joint probability of agreement, and given that high overall agreement is a desired feature of assessment measures, it is not completely unheard of to simply measure the reliability of a measure based on this statistic.

Furthermore, while it is important to measure inter-rater agreement at the item level for purposes of refining and validating the guidelines for human raters' application of the metrics, in practice, values produced by the metrics will be examined and compared at the level of entire texts or text collections, and not at the level of individual sentences. Thus, it will be important to assess inter-rater reliability for the metrics at this higher level – a level at which minor differences between raters on single items will have less importance. This introduces an interesting twist to the computation of inter-rater reliability: While it is possible – although possibly not ideal – to use kappa to measure inter-rater agreement at the item level, where the judgements can be seen as categorical, the aggregate values of these metrics at the text level results in a continuous variable. Thus, it will be necessary to use alternate methods to measure agreement at that level.

Finally, not to lose sight of the overall goal of our research program, we will strive to develop linguistically-based evaluation metrics that are interpretable, and exhibit high inter-rater agreement. We will continue to explore correlations between these PLATO metrics and more efficiently computable automated and semi-automated metrics, and also between the PLATO metrics and information processing tasks that humans and automated downstream processes are able to perform on MT output.

7 Acknowledgements

We acknowledge and thank Lacey Sculley, who spent many long hours learning the PLATO assessment suite, developing materials, and training the assessors, Jonathan Phillips and David DeBarr of MITRE, Ann Brodeen of ARL, and last, but certainly not least, the 20 PLATO assessors who joined us in this investigation.

References

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition. New York: Wiley.

Babych, B. and A. Hartley (2004) *Extending the BLEU Evaluation Method with Frequency Weightings*. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the kappa statistic. *Computational Linguistics* 22(2): 249-254.

Church, K. and E. Hovy. 1993. Good applications for Crummy Machine Translation. *Machine Translation* 8:239-258.

Doyon, J., Taylor, K., and J. White. 1999. Task-Based Evaluation for Machine Translation. *Proceedings of MT Summit 7*. Singapore.

Hovy, E. 1999. Toward Finely Differentiated Evaluation Metrics for Machine Translation. *Proceedings of the EAGLES Workshop on Standards and Evaluation*. Pisa, Italy.

International Standards for Language Engineering. 2000. (http://www.isi.edu/natural-language/mteval) The ISLE Classification of Machine Translation Evaluations, Draft 1, October, 2000. *Proceedings of the Hands-on Workshop on Machine Translation Evaluation. Association for Machine Translation in the Americas*, Cuernavaca, Mexico.

Jones, D. and G. Rusk. 2000. *Toward a Scoring Function for Quality-Driven Machine Translation*. In Proceedings of COLING-2000.

Lavie et al. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In Machine Translaiton: From Real Users to Research, *Proceedings of the Sixth Conference of the AMTA*, Washington, D.C.

Lin, C-Y., F. Och (2004*) Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics*. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain.

Mann, W., and S. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8:3.243-281.

Melamed, I, R. Green, and J. Turian (2003) *Precision and Recall of Machine Translation*. In Proceedings on NAACL/HLT, Edmonton, Canada.

Miller, Keith (2004). *PLATO: Predictive Linguistic Assessment of Translation Output*. Panel: Overview of Current Trends in MT Evaluation. The 6th Conference of the Association for Machine Translation in the Americas. Georgetown University, Washington, DC.

Miller, K. and M. Vanni. 2001. Scaling the ISLE Taxonomy : Development of Metrics for the Multi-Dimensional Measurement of Machine Translation Quality. *Proceedings of MT Summit VIII*. Santiago de Campostela, Spain.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2001. BLEU : *A Method for Automatic Evaluation of Machine Translation*, IBM Research Report, RC22176, September 2001.

Papineni, K., S. Roukos, T. Ward, J. Henderson, and F. Reeder. 2002. Corpus-Based Comprehensive and Diagnostic MT Evaluation : Initial Arabic, Chinese, French, and Spanish Results. *Proceedings of the Human Language Technology Conference*.

Taylor, K. and J. White. 1998. Predicting What MT is Good for : User Judgments and Task Performance. *Proceedings of the 1998 conference of the Association of Machine Translation in the Americas*. 364-373.

Taylor, K. and J. White. (1998). *Predicting What MT is Good For*. EAGLES Workshop. Geneva, Switzerland.

Uebersax, J. Statistical Methods for Rater Agreement. 2003. http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm

Vanni, M. and K. Miller. 2002. Scaling the ISLE Framework : Use of Existing Corpus Resources for Validation of MT Evaluation Metrics Across Languages. *Proceedings of Language Resources and Evaluation Conference,* Spain.

White, John S. and K. Taylor. (1998). Task-Oriented Metric for MT Evaluation *Proceedings of AMTA 1998*, Langhorne, PA.

White, John S., Jennifer B. Doyon, and Susan W. Talbott. (2000). Task Tolerance of MT Output in Integrated Text Processes. *Proceedings of ANLP-NAACL 2000 Workshop on Embedded MT Systems.*

White, J.S. and T.A. O'Connell. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Further Approaches. *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*. Columbia, MD.

Weinberg, Amy (2005). Machine Translation Evaluation at the Center for Advanced Study of Language.. *Workshop on Intelligence Analysis,* Mitre Corporation.