# Translation Adequacy and Preference Evaluation Tool (TAP-ET)

**Mark Przybocki, Kay Peterson, Sébastien Bronsart**
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
E-mail: mark.przybocki@nist.gov, kay.peterson@nist.gov, sebastien.bronsart@nist.gov

### Abstract

Evaluation of Machine Translation (MT) technology is often tied to the requirement for tedious manual judgments of translation quality. While automated MT metrology continues to be an active area of research, a well known and often accepted standard metric is the manual human assessment of adequacy and fluency. There are several software packages (RWTH, 2000) (LDC, 2005) that have been used to facilitate these judgments, but for the 2008 NIST Open MT Evaluation (NIST, 2008), NIST's Speech Group created an online software tool to accommodate the requirement for centralized data and distributed judges. This paper introduces the NIST TAP-ET application and reviews the reasoning underlying its design. Where available, analysis of data sets judged for Adequacy and Preference using the TAP-ET application will be presented. TAP-ET is freely available and ready to download, and contains a variety of customizable features.

## 1. Introduction

NIST's Speech Group has coordinated annual evaluations of MT technology for text-to-text systems since 2002 (NIST, 2008). Our evaluation metric of choice has been BLEU (Papineni, 2002), in part because it is an automatic and repeatable metric which makes it usable for automatic machine learning techniques, but more importantly because BLEU has shown general correlation with human judgments of adequacy and fluency. Our evaluation practice has been to declare a single primary metric to which system developers may choose to optimize their system performance. When done, this results in a clearer picture of the strengths and weaknesses of different algorithmic approaches, regardless of the metric chosen as primary. For the NIST Open MT evaluations, BLEU is an approximation metric in that a consistent positive correlation between *improved* BLEU scores and general translation quality has been demonstrated. Each year, significant time was invested post-evaluation to have the system translations judged for adequacy (and sometimes fluency) by human assessors. The information gained from this process is used to validate our use of BLEU, and other automatic metrics, for ranking systems in evaluation. It also serves as an alternative metric for rule-based approaches, which many have found to be unfairly represented by several automated metrics. Unfortunately, the time and effort required for human assessments has meant that only a portion of the evaluation test set could be processed in a timely manner – typically about 10% of the evaluation test set, and for only a handful of systems.

The 2008 NIST Open MT evaluation (MT08) expanded in interesting directions, including the addition of new language pairs, a bi-directional test, a sequestered "Progress" test, and the inclusion of DLPT* comprehension tests (Jones, 2007). Human assessments were funded under all previous Open MT evaluations, but this was not the case for MT08. And since NIST realized the importance the role human assessments hold in MT evaluation, we supported a volunteer-based model similar to (Koehn, 2007) to generate human assessments of MT08 systems. To support this model, a new software tool, the "Translation Adequacy and Preference Evaluation Tool", or TAP-ET, was created, and in doing so, the implementation design used in previous NIST MT evaluations was completely re-evaluated.

## 2. Human Assessment Test Types

There are several types of human assessments of MT one might wish to implement. These include:

- Adequacy: How well does the translation match the reference(s) in meaning?
- Fluency: How natural is the resulting translation?
- Application-specific: Does the translation meet an application need (form filling, named entities, index and searching, …)?
- Preference: Given two versions of a translation, which is preferred?
- Odds of Successful Transfer of Low-level Concepts: How many low-level concepts present in the source language are represented in the target language? (Sanders, 2008)

We designed TAP-ET to investigate the collection of two of these types of assessments: *Adequacy* and *Preference* judgments. Future versions of the software may be configurable to include other forms of human assessments, such as *Fluency*.

### 2.1 Adequacy Testing

Human assessments of adequacy are the most popular and trusted manual measure of MT quality. The general implementation of this test consists of showing a judge one or more reference translation(s)[1] and a system translation of the same sentence. The judge makes a combined quantitative and qualitative decision as to how adequate the translation is compared to the reference. This score is often recorded as a point on a multipoint scale. Various implementations make use of different scales, but many employ either a 5-point or a 7-point scale. There has been recent work experimenting with a continuous scale (Mathieson, 2003) for web surveys. This approach was briefly considered for TAP-ET, but the inter-judge agreement lagged the performance of the current implementation.

The Linguistic Data Consortium generated the human assessments for the 2002-2006 NIST Open MT evaluations using a 5-point scale and asking the following question:

How much of the meaning expressed in the Reference translation is also expressed in the System translation?

_ All _ Most _ Much _ Little _ None

Virtually no instructions or examples were given to assist the judges in making their decisions. Judges were qualified individuals with an academic background in linguistics with good understanding of what was being measured.

We identified several weaknesses with the previous implementation, some of which in include:

- The choice difference between two adjacent categories can be confusing, sometimes due to the anchor point descriptors and sometimes due to insufficient guidance and examples. For example, the semantic difference between "**Most**" and "**Much**" is unclear.
- There was insufficient guidance on what issues this question really attempts to measure. Should only the presence of information from the reference count towards the score? If so, the introduction of misleading information as well as overall meaning transfer are not adequately reflected.
- It may be tempting for judges to continually select the same category for many consecutive decisions, without the necessary reflection.
- The interface lacked a visual guidance mechanism to aid the judges in comparing system to reference translation.

To address these weaknesses, we redesigned the adequacy measure for MT08, separating it into a more quantitative and more qualitative decision process. After careful investigation and consultation with expert users and implementers of human assessment software, the following framework was chosen. The first part of the new adequacy judgment was worded the same, but the descriptors of the scale points changed to:

How much of the meaning expressed in the Reference translation is also expressed in the System translation?

_ All _ Much _ Half _ Little _ None

Guidelines accompanying this quantitative question made it clear that the judgment should pertain to only the presence of information existing in the reference translation. The anchor points were renamed to make the difference between scale points more consistent. The guidelines contained carefully selected examples of each anchor point.

A machine translation may contain a lot, or even all, the concepts of the reference translation, but there are cases in which the main meaning of the sentence is still not conveyed. This can be due to ordering problems, the introduction of misleading information, or mistranslation of absolutely vital pieces of meaning such as polarity. We believe that measuring the absence/presence of the "essential meaning" is part of obtaining a complete description of translation adequacy.

We introduced an additional question as we realized that a one-dimensional scale alone cannot handle the task of determining adequacy in a fully satisfactory way.

When a translation received one of the top three scores, a second question of a more global qualitative nature was asked:

Does the System translation mean essentially the same as the Reference translation?

__ Yes __ No

The second question was designed to help distinguish between the higher performing systems. Also, it may help address the danger of assessments being made too hastily without sufficient reflection. In the end, this single question may prove to be the best determiner of system quality, as it is focused on the purpose of the MT system.

This initial design was tested using system translations collected during the 2006 NIST Open MT evaluation. Through testing, it was revealed that judges often selected a common default score of "**Much/No**" and indicated that they were finding many of the system translations to contain some level more than "**Half**" but certainly not "**All**" of the information. This finding is due in part to the

[1] A reference translation is a high quality manual translation of the source data by one or more bilingual speaker(s) of both the source and target language. Reference translations are used by BLEU to score system translations against, and allow for human assessments of source and target without the need for bilingual judges.

quality of current state-of-the-art of MT technology and was a sign that a more fine-grained separation of scores would be required to differentiate system performance at current capabilities.

Midway through our testing, we updated TAP-ET to use a 7-point scale with the following anchor points:

_ [7]-All _ [6] _ [5] _ [4]-Half _ [3] _ [2] _ [1]-None

[The numbers in square brackets are NOT displayed to the judge; they are included in this paper as reference points.]

Note the two unlabeled fields between "**Half**" and the two extremes. Points **[5]** and **[6]** allow for judges to express the "**Much-No**" common cases mentioned above, in finer detail, where their decisions may now lean towards the middle or extreme score. These are the cases that were found to be the most difficult to separate. The "Yes/No" question was now only applied to the top three categories, **[5], [6],** and **[7].**

On completion of our testing, each segment had received two scores, one using our original 5-point scale, the second from our 7-point scale.

As we continued to test TAP-ET, we found that judges were making good use of these intermediate points of the 7-point scale. While 43% of all decisions made using the 5-point scale fell into the "**Much**" category, 53% of all decisions made using the 7-point scale fell into either the **[5]** or **[6]** category, with 15% for **[5]** and 38.0% for **[6]**. The 10% increase is understandable, as judges now have finer grained decision points for what may have been "**Half**" and "**All**" decisions on the 5-point scale.

The weakness of insufficient visual guidance was addressed by building on a concept presented by (Voss, 2006) at the NIST MT06 evaluation workshop, where she discussed various techniques for providing feedback to judges. TAP-ET implements a simple shading scheme that identifies word matches between the system translation and the reference(s). This may aid in the consistency of the judgments by differentiating between phrase matches that are full of content words, and those with equal matches but of less importance to the meaning.

To achieve perfect inter-judge agreement is likely impossible due to the complexity and subjective nature of the task and the many different styles of machine translation output. But we anticipated that the more rigorous design, a new scale and descriptors, and improved guidelines with carefully selected examples, would result in an acceptable level of inter-judge agreement while providing more detailed information for analysis. This will be explored below in the *Results* section, as well as during our presentation at the 2008 Language Resources and Evaluation Conference (LREC).

### 2.2 Preference Testing

The second method of human assessment implemented in TAP-ET was pair-wise comparisons of different translations, referred to as *Preference* assessment. Preference judgments were not included in previous NIST MT evaluations. The design is quite simple; judges see one reference and two versions of system translations. They then choose between three decision points:

__ Prefer System #1 __ No Preference __ Prefer System #2

In this type of test, it is important to display translations for a specific system randomly so that they sometimes appear as "System #1" and sometimes as "System #2".

Judges found the process for selecting preference to be very tedious. This was partly due to the size of our test corpus, which contained 25 documents totaling 249 segments for each of eight system translations. For full preference determination of the eight systems, 28 comparisons were required for each segment. In an effort to make this a manageable task, preference judgments were limited to the first four segments of each document. Also, a single segment was assessed for all 28 pair-wise combinations in a row, randomly selecting the order in which the system pairs were presented. By assessing a single segment in this manner, we reduced the burden of re-learning the reference translation. We presented succeeding segments in document order to maintain context.

## 3. Volunteer Based Assessment Model

For MT08, human assessments were implemented using a volunteer based model. Assessments of *Adequacy* and *Preference* were limited to system translations submitted by participants who provided volunteer judges. Judges were required to be native or near-native speakers of English. Their task was to perform assessments on a data set that was expected to take approximately ten hours to complete. Part of their own site's data was blindly included in their assignment.

There were 34 independent researcher teams participating in MT08 and an additional 6 teams working in some form of collaboration. Of the 40 participants, 50% signed up for the human assessments. Some of them provided multiple judges, which allowed them to have multiple systems included in the assessment process. The pool of participant judges was rounded out by 13 volunteers from non-participant affiliations.

## 4. TAP-ET Implementation

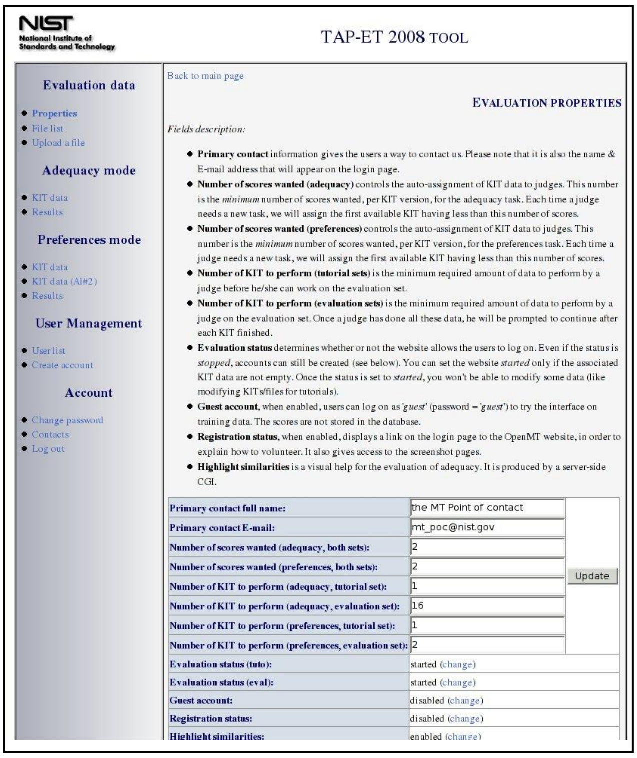
NIST implemented TAP-ET in a PHP/MySQL application. The software was built with the MT08 evaluation in mind, but it includes flexibility that will make it useful outside of the NIST evaluations. TAP-ET is freely distributable (NIST Tools, 2008).

Access to the TAP-ET application is usually password-protected in order to restrict access to those participating in the assessment task.  More importantly, this provides a mechanism of tracking the progress of each judge, necessary since some “volunteers” are less enthusiastic than others in completing their assignments, and may need a reminder.  Guest accounts are available.

The TAP-ET application has two main sections; the “Administrative Interface,” which controls setting up the evaluation data to be assessed, and the “Judge Interface,” which contains instructions as well as access to a judge’s tasks.

## 4.1 Administrative Interface

This section is only accessible to accounts bearing special privileges, usually owned by the evaluation coordinator(s). The setup of an evaluation is accomplished in three steps:



*Fields description:*

- **Primary contact** information gives the users a way to contact us. Please note that it is also the name & E-mail address that will appear on the login page.
- **Number of scores wanted (adequacy)** controls the auto-assignment of KIT data to judges. This number is the *minimum* number of scores wanted, per KIT version, for the adequacy task. Each time a judge needs a new task, we will assign the first available KIT having less than this number of scores.
- **Number of scores wanted (preferences)** controls the auto-assignment of KIT data to judges. This number is the *minimum* number of scores wanted, per KIT version, for the preferences task. Each time a judge needs a new task, we will assign the first available KIT having less than this number of scores.
- **Number of KIT to perform (tutorial sets)** is the minimum required amount of data to perform by a judge before he/she can work on the evaluation set.
- **Number of KIT to perform (evaluation sets)** is the minimum required amount of data to perform by a judge on the evaluation set. Once a judge has done all these data, he will be prompted to continue after each KIT finished.
- **Evaluation status** determines whether or not the website allows the users to log on. Even if the status is *stopped*, accounts can still be created (see below). You can set the website *started* only if the associated KIT data are not empty. Once the status is set to *started*, you won't be able to modify some data (like modifying KITs/files for tutorials).
- **Guest account**, when enabled, users can log on as '*guest*' (password = '*guest*') to try the interface on training data. The scores are not stored in the database.
- **Registration status**, when enabled, displays a link on the login page to the OpenMT website, in order to explain how to volunteer. It also gives access to the screenshot pages.
- **Highlight similarities** is a visual help for the evaluation of adequacy. It is produced by a server-side CGI.

| | |
|---|---|
| **Primary contact full name:** | the MT Point of contact |
| **Primary contact E-mail:** | mt_poc@nist.gov |
| **Number of scores wanted (adequacy, both sets):** | 2 |
| **Number of scores wanted (preferences, both sets):** | 2 |
| **Number of KIT to perform (adequacy, tutorial set):** | 1 |
| **Number of KIT to perform (adequacy, evaluation set):** | 16 |
| **Number of KIT to perform (preferences, tutorial set):** | 1 |
| **Number of KIT to perform (preferences, evaluation set):** | 2 |
| **Evaluation status (tuto):** | started (change) |
| **Evaluation status (eval):** | started (change) |
| **Guest account:** | disabled (change) |
| **Registration status:** | disabled (change) |
| **Highlight similarities:** | enabled (change) |

**Figure 1**: Screen shot of a sample page from the administrative interface of TAP-ET.

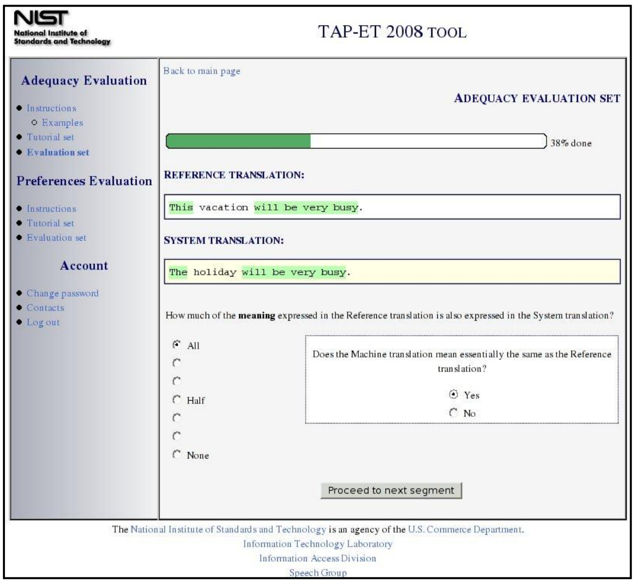
The administrator performs the first step offline by identifying the system translations to be included in the assessments.  Often, less than the entire evaluation set is to be assessed, in which case a document filtering operation and a system filtering operation are necessary. Optionally, a tutorial set may be identified for use in training new judges before they process the evaluation data.  Both the tutorial and the evaluation data set contain several system files and one or more reference files. Special care must be taken to ensure that the files are consistent (e.g., for number of documents, segments, identifying attributes).  Once identified and verified, the files may be uploaded to the application, which takes care of populating the corresponding SQL tables.

Second, a dedicated interface allows defining “kits”, packets of data that are presented to each judge. The interface asks several questions to structure how and what data is included in each kit (e.g., “How many documents per kit?”).  Once all the questions are answered, the application creates the kits automatically.

Third, the administrator sets the final evaluation parameters and finalizes the user login information.  The most important parameters concern the assignment of kits to judges, determining how many judgments are expected per kit, as well as the total number of kits each judge will process.

## 4.2 Judges Interface

Once an evaluation is set up, judges receive their log-in PIN and password. For MT08, PINs were distributed in a way that did not require NIST to record personal information about the judges.  After successfully accessing the system, judges are presented with a brief summary of their remaining tasks.



REFERENCE TRANSLATION:

This vacation will be very busy.

SYSTEM TRANSLATION:

The holiday will be very busy.

How much of the **meaning** expressed in the Reference translation is also expressed in the System translation?

● All
○
○
○ Half
○
○
○ None

Does the Machine translation mean essentially the same as the Reference translation?

● Yes
○ No

The National Institute of Standards and Technology is an agency of the U.S. Commerce Department.
Information Technology Laboratory
Information Access Division
Speech Group

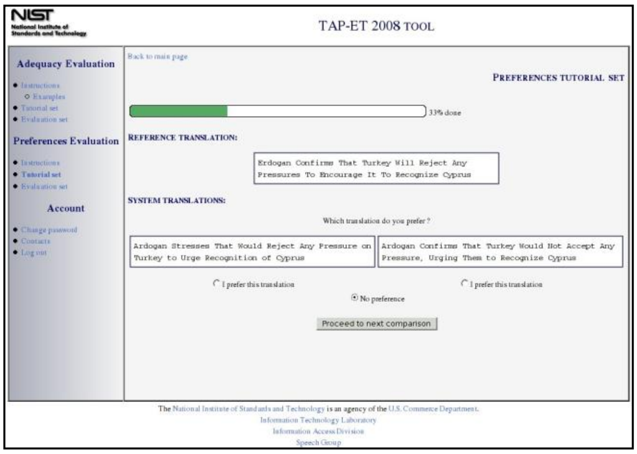
**Figure 2**: Screen shot of the Adequacy judgment screen.

The tasks each judge must complete include:

- Adequacy tutorial data
- Adequacy evaluation data
- Preference tutorial data
- Preference evaluation data

The choice of tasks is subject to these restrictions:

1) The tutorial tasks must be completed (only once) before starting the evaluation data.
2) The Adequacy tasks must be completed before starting the Preference tasks.



PREFERENCES TUTORIAL SET

REFERENCE TRANSLATION:

Erdogan Confirms That Turkey Will Reject Any Pressures To Encourage It To Recognize Cyprus

SYSTEM TRANSLATIONS:

Which translation do you prefer ?

| | |
|---|---|
| Ardogan Stresses That Would Reject Any Pressure on Turkey to Urge Recognition of Cyprus | Ardogan Confirms That Turkey Would Not Accept Any Pressure, Urging Them to Recognize Cyprus |
| ○ I prefer this translation | ○ I prefer this translation |

● No preference

The National Institute of Standards and Technology is an agency of the U.S. Commerce Department.
Information Technology Laboratory
Information Access Division
Speech Group

**Figure 3**: Screen shot of the Preference judgment screen.

The tutorial part gives judges a small set of data on which to (1) become accustomed to the TAP-ET interface, and (2) self-calibrate themselves before working on the evaluation data.

Users may log off at any time and on return, the application picks up where they left off. The instructions for the tasks, along with examples of the different adequacy scores, are available at http://ww.nist.gov/speech/tests/mt/2008/ha.

TAP-ET displays tasks using a standard HTML form. Once a judge makes a decision for a segment, the results are processed by PHP (server-side) and inserted into the database. TAP-ET records time markers to obtain statistics regarding the length of time it takes to make each decision. Judges' decisions are final; it is not possible for a judge to go back to change a previous answer.

To help identify (and correct) possible assessment errors, the NIST-internal version of TAP-ET includes a customized interface that allows for easy adjudication over pairs of judgments. Future public versions of TAP-ET may include this option.

## 5. Results

An important goal for the design of TAP-ET was to create an online application that would support collecting consistent and useful human assessments with minimal burden to the user.

We report results from two large-scale usages of TAP-ET that provide feedback for future application improvements. We examine the inter-judge agreement for Adequacy and Preference assessments and explore the correlation between the two. We also review feedback from judges for possible TAP-ET improvements.

Section 5.1 describes the first large-scale usage, a test corpus used to design, build, and test TAP-ET. The second large-scale usage, the MT08 evaluation for which TAP-ET was designed, is described in section 5.2.

### 5.1 TAP-ET Testing Corpus

NIST selected system translations from the MT06 evaluation and from a handful of TRANSTAC training dialogs to test the TAP-ET application. See Table 1 for the test corpus statistics.

### 5.1.1 Adequacy Judgments on TCC

There were three judges assessing adequacy using the TAP-ET Testing Corpus (TTC). Judges assessed equal amounts of data, such that each segment received two adequacy scores. Every segment was first judged using the newly designed 5-point scale as described above. The second score came from the 7-point scale currently implemented in TAP-ET.

| TAP-ET Testing Corpus (TTC) Statistics | |
|---|---|
| Data Set | MT-06 |
| Genre | Newswire |
| Number of documents | 25 |
| Total number of segments | 249 |
| Source Language | Arabic |
| Target Language | English |
| Number of system translations | 8 |
| Data Set | TRANSTAC |
| Genre | training dialogs |
| Number of documents | 1 |
| Total number of segments | 16 |
| Source Language | Iraqi Arabic |
| Target Language | English |
| Number of system translations | 5 |

**Table 1:** TTC data used in original testing of the TAP-ET application.

Adequacy was performed over the entire TTC, while Preference was limited to the first four segments of each document in an effort to make the test more manageable.

Figure 4 shows the inter-judge agreement rates achieved for the past three NIST Open MT evaluations as well as two types of comparisons obtained using the Newswire data from TTC. For each test set, we show two types of agreement rate, "exact match", defined as two judges assigning the exact score, and "1 category off", defined as two judges assigning a score that is equal or in adjacent categories. The qualitative "Yes/No" question is not used for this analysis.

The two categories for TTC assessments include:

1. "Mapped 5pt" - mapping of the 7-point scale into the corresponding 5-point scale.
2. "7-pt Czar" – compares the judgment using the 7 point scale to the adjudicated decision.

The exact match rate ranges between 34-46% for the previous NIST Open MT evaluations. Using the TAP-ET application and associated guidelines, we found exact matches of 55% when we limit analysis to the data assessed using the 7-point scale mapped back to the 5-point scale for comparison. The exact match between the adjudicated score (on the 7-point scale) and that of the original 7-point scale decision was 62%.

When we looked at the 1-off category, the range of agreement for the NIST Open MT evaluations was 80-90%, while the two categories of the TTC were consistent at 93% agreement.

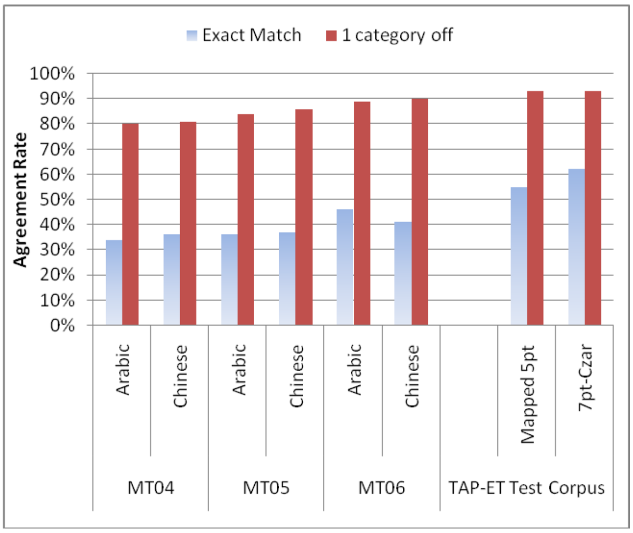


**Figure 4:** Inter-judge Adequacy agreement rates, limiting the data to Newswire for better comparison. Two judges independently scored each test segment.

The TRANSTAC data was evaluated separately and only using the 7-point scale. The inter-judge Adequacy agreement rate was 55%. We note here that forty percent of the segments received three scores in this calculation. The agreement at the 1-off category was 85%. Note that we do not have any other comparable data sets for multiple judgments using a 7-point scale. The TRANSTAC data was generally shorter and much less complex than the other data assessed during testing; we believe this greatly influenced the agreement rates.

While we are encouraged that our findings show improvement over results obtained in the past, we realize that the Adequacy assessments created for the development of TAP-ET came from judges who were more in tune with the application than the volunteers we expect to recruit for future efforts.

Instructions to judges of past Open MT evaluations of human assessment encouraged the decisions to be made in "less than 30 seconds". We did not use such a time constraint here. Table 2 lists estimates (based on averages) of time required by the judges to make Adequacy decisions on the TTC data.

| Time Requirements: Adequacy Judgments | | |
|---|---|---|
| Data Set | Estimated Time | Average Time |
| MT06 | 30 seconds | 47 seconds |
| TRANSTAC | 30 seconds | 36 seconds |

**Table 2:** Average time spent on human assessment of Adequacy, limited to decisions made in two minutes or less (over two minutes may indicate a pause in work).

### 5.1.2 Preference Judgments on TCC

Preference judgments were completed for the TRANSTAC dialog segments included in the TCC. These segments were rather short, less than 20 words on average. Three judges provided scores for each preference comparison.

The three-way inter-judge agreement for this data was 58%, with most disagreements involving one judge selecting "No preference" while the others had a preference. When removing these "No preference" decisions, the agreement rate between the three judges rises to 90%.

Similar analysis for the MT06 TTC data shows a disagreement rate between two independent judgments of preference at 4.2% (33 disagreements out of 784 samples). This is again with "No preference" removed from analysis.

Table 3 lists estimates of time required by judges to make the Preference decisions on the TTC data.

| Time Requirements: Preference Judgments | | |
|---|---|---|
| Data Set | Estimated Time | Actual Time |
| MT06 | 30 seconds | 25 seconds |
| TRANSTAC | 30 seconds | 19 seconds |

**Table 3:** Average time spent on human assessment of Preference, limited to decisions made in two minutes or less (over two minutes may indicate a pause in work).

### 5.1.3 Correlation between Adequacy and Preference

Adequacy and Preference are two different types of manual assessments which are both considered a means for determining quality of machine translations. In this section, we analyze scores on segments that received both types of judgments and were adjudicated to one score. We examine if the two types of tests agree.

We claim that the two tests agree if system-A's segment is chosen as being preferred over system-B's segment in the Preference test, and system-A's segment received an equal or higher score than system-B's segment in the Adequacy test. Otherwise, the two tests identify opposite systems as being the better system.

The presence of the "Yes/No" data point for determining essential meaning allows for several methods of ranking Adequacy scores. For the following analysis, we rank the

Adequacy scores as follows (high to low): **[7]-Yes, [7]-No, [6]-Yes, [6]-No, [5]-Yes, [5]-No, [4], [3], [2], [1]**.

In 40 samples of TRANSTAC TTC data, there were four instances where the Preference judgment did not match the Adequacy score assigned. Two of the mismatches occurred when comparing segments with Adequacy scores **[1]** and **[2]** and the adjudicated Preference score was "No preference". The third mismatch occurred for segments with Adequacy scores **[7]-No** and **[6]-yes** and the adjudicated Preference judgment was for the **[6]-yes** segment. The forth mismatch occurred for segments with Adequacy scores **[6]-yes** and **[6]-no** and the adjudicated Preference score was again "No preference".

In 370 samples of the MT06 TTC data, there were 26 instances (7%) where the Preference judgment did not match the Adequacy score assigned. Half of these mismatch cases occurred when the comparison of translations was within one category for Adequacy and the adjudicated Preference was "No preference".

## 5.2 NIST Open MT 2008 Evaluation Translations

TAP-ET was designed for the volunteer model of human assessments employed in MT08. At the time of paper submission, the assessments for MT08 were completed, but the analysis was not. We report preliminary results here, with more details planned for our presentation at LREC 2008. See Table 4 for MT08 assessment data statistics.

| MT08 Human Assessment Corpus | | | |
|---|---|---|---|
| | Arabic | Chinese | Urdu |
| Genre | Newswire & Web | Newswire & Web | Newswire & Web |
| Num. of documents | 26 | 23 | 26 |
| Num. of segments | 213 | 204 | 217 |
| Num. of systems | 12 | 15 | 6 |
| Num. of judges | 21 | 23 | 9 |

**Table 4:** MT08 data used for volunteer assessments.

### 5.2.1 Adequacy Judgments on MT08

Assessments of Adequacy were performed on three of the four language pairs offered for MT08. The high participation rate resulted in a larger pool of judges than was used in previous Open MT assessment exercises. All adequacy scores were based on the 7-point scale, and each segment for each system received two independent assessments. The inter-judge "exact match" rates for the two scores are:

- Arabic, Range: 20-38%, Average: 31%
- Chinese, Range: 23-45%, Average: 31%
- Urdu, Range: 23-32%, Average: 28%

The "1-category off" scores are:

- Arabic, Range 51-82%, Average: 70%
- Chinese, Range: 52-83%, Average: 70%
- Urdu, Range: 55-69%, Average: 63%

These numbers are slightly lower than what we observed in the testing of the application, and lower than what was achieved in past Open MT evaluations. There are two possible explanations for this. First, previous analyses involved assessments that in part (or in whole) used a five point scale, not the finer grained scale used here which may show benefit in differentiating between system performances. Second, some of our volunteer judges may not have been qualified as a judge, and consistency may have been a factor. Through an adjudication step, we will identify any such judges. Analysis is ongoing.

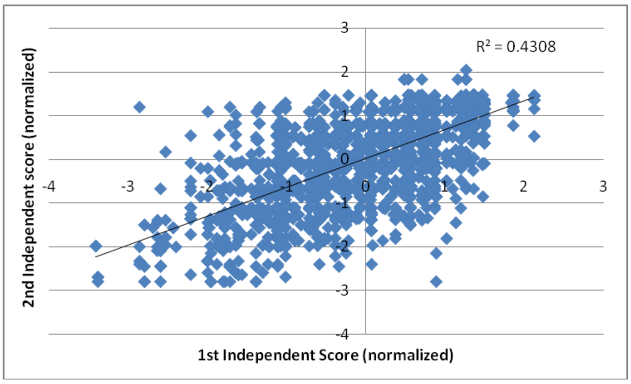


**Figure 5**: Scatter plot of normalized judge scores for the Arabic data assessed in MT08.

In Figure 5, we show the agreement of adequacy scores as a scatter plot. We normalized the scores according to the following formula:

$$\text{Score}_{norm}(\text{segment,judge})=\frac{\text{Score(segment,judge)-Mean(judge)}}{\text{StdDev(judge)}}$$

The R-squared value of 43% shows the level of correlation between the two judges. In Figure 6, we show the agreement rates as a function of category distance for each of the three languages assessed. While the "exact" match is low for the 7-point scale, the agreement rate quickly rises in the 1-off and 2-off categories.

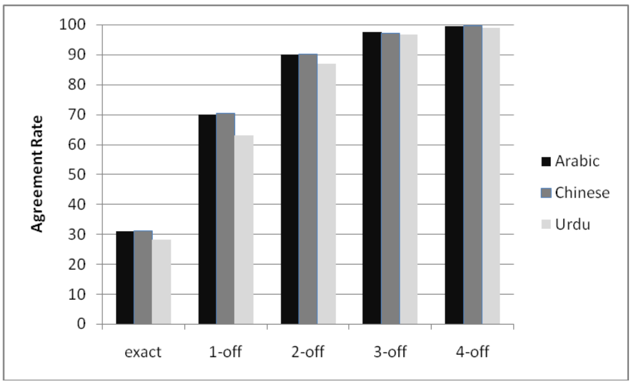


**Figure 6:** Agreement rates for adequacy shown by category distance between judges.

### 5.2.2 User Feedback on MT08

After finishing the MT08 assessments, we gathered feedback from the volunteer judges. Overall, the TAP-ET application was well received. But, as we had expected given that MT08 was our first "real world" test of the tool, we also received numerous suggestions for improvements. We list the most important ones here.

For the Preference assessments, we will abandon presenting individual pair-wise comparisons in favor of displaying all system translations of a given segment at once. TAP-ET will allow the judge to re-order the translations to the ranking of their preference. This will mean a dramatic reduction in the number of judgments needed to be made and (hopefully) to the time required in making them.

Also for the Preference assessments, we will investigate replacing the "No preference" button with two buttons, one that will identify when both translations are equally good, and one that will identify when both translations equally bad.

We will improve the guidelines to give guidance on the use of such aspects as world knowledge, outside knowledge, and context for making judgments.

We will investigate several options for improving the usability of TAP-ET, for example by making the clickable areas larger and by adding keyboard shortcuts as an alternative to using the mouse.

Several judges reporting taking more time than we had estimated per judgment; we will recalibrate our estimations based on the actual usage as tallied for MT08.

## 6. Summary

This paper introduces the first version of TAP-ET, a freely downloadable application for use in generating MT human assessments of Adequacy and Preference. It meets the need of many MT evaluations for centralized data and administration but distributed judges and thus will support future evaluations at NIST and elsewhere.

Preliminary results using a test corpus have demonstrated similar rates of inter-judge agreement as were achieved in previous NIST Open MT evaluations. Results from the effort the application was designed to support, the NIST Open MT08 evaluation, are preliminary but they, too, show acceptable rates of agreement.

The introduction of Preference judgments as a complementary type of human assessment will enable alternative methods for determining system ranking and provides a valuable error analysis tool for system developers.

Much effort went into the design of TAP-ET, including the creation of new scales for Adequacy, a detailed set of examples for each Adequacy anchor point, and improved guidelines. User feedback and analysis of results will be used to further refine the application as we strive to improve inter-judge agreement and to reduce the burden placed on humans performing these necessary judgments.

## 7. Disclaimer

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

## 8. References

Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., and Emonts, M., (2007). ILR-Based MT Comprehension Tests with Multi-Level Questions. Human Language Technology Conference, 23 April 2007, Rochester, NY (http://www.mit.edu/~dajones)

Koehn, P., (2007). Evaluating evaluation Lessons from the WMT 2007 Shared Task, http://www.elra.info/mtsummit2007/Pres-2-Koehn.pdf

LDC (Linguistic Data Consortium) (2005). *Linguistic Data Annotation Specification Assessment of Fluency and Adequacy in Translations*, http://projects.ldc.upenn.edu/TIDES/tidesmt.html

Mathieson, K., and Doane, D., (2003). Using Fine-Grained Likert Scales in Web Surveys. *Alliance Journal of Business Research*, 1 (1), April 2005, pp. 27-–34.

NIST (National Institute of Standards and Technology) (2008). *NIST 2008 Open MT Evaluation*, http://www.nist.gov/speech/tests/mt/2008

NIST Tools, Speech Group Evaluation Software Tools (2008). http://www.nist.gov/speech/tools

Papineni, K., Roukos, S., Ward, T., and Zhu, W.J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In the proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (*ACL-2002*), pp. 311–318.

RWTH Aachen (2000). *EvalTrans – Fast Evaluation of MT Research*, http://www-i6.informatik.rwth-aachen.de/web/Software/EvalTrans

Sanders, G., Bronsart, S., Condon, S., and Schlenoff, G. (2008). Odds of Successful Transfer of Low-level Concepts: A Key Metric for Bidirectional Speech-to-speech Machine Translation in DARPA's TRANSTAC Program. In the proceedings of the sixth international conference on Language Resources and Evaluation, LREC 2008.

Voss, C., (2006). Highlighter Experiment, *NIST Open MT-2006 workshop*.