Task-based MT Evaluation: Tackling Software, Experimental Design, & Statistical Models

Calandra Tate [1, 3]
ctate@math.umd.edu
[1] Dept. of Mathematics
University of Maryland
College Park, MD 20740

Sooyon Lee[2 3]
sylee@arl.army.mil
[2]ARTI, Inc.
Alexandria, VA 22314

Clare R. Voss[3]
voss@arl.army.mil
[3]Multilingual ComputingGroup
Army Research Laboratory
Adelphi, MD 20783

Abstract

Even with recent, renewed attention to MT evaluation—due in part to n-gram-based metrics (Papineni et al., 2001; Doddington, 2002) and the extensive, online catalogue of MT metrics on the ISLE project (Hovy et al., 2001, 2003), few reports involving task-based metrics have surfaced. This paper presents our work on three parts of task-based MT evaluation: (i) software to track and record users' task performance via a browser, run from a desktop computer or remotely over the web, (ii) factorial experimental design with replicate observations to compare the MT engines, based on the accuracy of users' task responses, and (iii) the use of chi-squared and generalized linear models (GLMs) to permit finer-grained data analyses. We report on the experimental results of a six-way document categorization task, used for the evaluation of three Korean-English MT engines. The statistical models of the probabilities of correct responses yield an ordering of the MT engines, with one engine having a statistically significant lead over the other two. Future research will involve testing user performance on linguistically more complex tasks, as well as extending our initial GLMs with the documents' Bleu scores as variables, to test the scores as independent predictors of task results.

1 Introduction

Given that MT systems are now accessible to a wide range of users via web-based search engines and low-cost software packages, the general question of which *actual tasks* can be carried out reliably with which MT systems now arises regularly among the users of these systems. This is especially true of users in business and in government who, for their work, require translation assistance with foreign language documents that they cannot read. They are wary of the accuracy of MT systems when they see output that is disfluent or incomplete. Google's decision to make MT available to their users, however, suggests that this wariness does not deter all users. Indeed it may be that Google has observed empirically what Levin et al. (2000) report in evaluating the even noisier case of speech translation: users of MT output perform their tasks at a much higher rate of success than would be expected, given the accuracy of the MT output.

For MT researchers, their attention to actual, practical tasks that even weak MT engines can perform, was first sparked by Church and Hovy (1993). A few years later, Taylor and White (1998) hypothesized that MT engines could be evaluated *operationally* in terms of a hierarchy of tasks. They proposed that the "weakest" engines be identified as those that only enable their users to perform the least linguistically demanding task in the hierarchy, while the "stronger" engines would be identified as those that enable their users to perform more linguistically demanding tasks. To the best of our knowledge, this was the first proposal for systematic, linguistically motivated, task-based MT evaluation.

Since then, some MT researchers have begun to examine evaluation tasks and report on these at workshops,[1] while others have focused on non-task-based, n-gram metrics (Papineni et al., 2001; Doddington, 2002 on DARPA TIDES[2] MTEval).[3] Curiously, there appear to be only a few researchers who both work with actual, current

[1] See Hovy et al. (2003) for a listing of website links to five such workshops since 2000.
[2] TIDES is a research program administered by DARPA, one agency within the US Department of Defense that has funded natural language processing research for many years.
[3] Indeed this interest in n-gram-based evaluation has extended beyond MT to research on summarization (Pastra et al. 2002) and headline generation (Zajic et al. 2002).

operational MT systems under development and conduct task-based user studies or experiments for evaluating those systems, that they then report on.[4] Since our users have specific tasks for which MT is needed, our goal has been to provide MT evaluation results to them in terms of their tasks. For them, a BLEU score of 0.35 or a NIST n-gram score of 7 will only become informative when such numbers correlate with the tasks they perform.[5]

This paper introduces our work on three parts of task-based MT evaluation, as developed and applied in rapidly assessing several MT engines: (i) the software that we designed, tested, and used to run experiments, where users' online actions, timing data, and task decisions are recorded automatically as users perform tasks via a browser, locally on desktop or remotely over the web, (ii) the factorial experimental design with replicate observations to compare the MT engines, based on the accuracy of users' task responses, and (iii) the use of chi-squared and generalized linear models (GLMs) to permit finer-grained data analyses. After a brief overview of our earlier work in section 2, the paper describes these three parts of our work in separate sections. The paper concludes with a preliminary analysis for our future work, comparing BLEU-based results with task-based results, and a summary of our experimental results.

2 Background

Our first step was to pilot a monolingual version of the document categorization task, with 12 English speakers judging 18 English documents, in order to establish the experimental procedure, the time requirements, an instruction set, and the level of accuracy at which the subjects could categorize the texts that we culled from the web, given a set of nine categories that we hypothesized would be intuitively clear. The subjects in this pilot performed the task with pen and paper. We found that they had no difficulty with the task, but some of the categories confused to them.

The next phase of our work involved twelve (different) English speakers who categorized 18 documents machine-translated from Korean into English. The test collection was constructed by a native Korean speaker who selected Korean news articles from the web and assigned them each to one of nine categories (two of which were changed after the monolingual pilot). The online articles were machine-translated twice, once by each of the two Korean-to-English MT engines to generate the test documents. This pilot was conducted with all subjects in the same room at once, hearing the same instructions together. The subjects each saw a different randomized presentation of an equal number of documents output by each system in each category. They never saw more than one translation of a particular source document. The subjects, who had no knowledge of Korean, performed the task with pen and paper and completed it in about half an hour. During an informal exit poll when asked about the task, most subjects said it was easy to understand.

After these two pilots—to speed up the process of creating randomized document sets for presentation, to reduce the time needed to code subjects' responses prior to data analyses, and to collect subject response times—we shifted from pen and paper to constructing our own software program to run the experiments. Last year, we conducted a third pilot on laptop computers, following the same procedure as before, but with documents translated by two different MT engines from Arabic, rather than Korean. Working on laptops, twelve subjects entered their task responses online, while the new software tracked their progress through pre-established, randomized sequences of 18 translated documents. While the subjects also reported that the task was easy to understand and we were pleased to be able to quickly run our data analyses with the online results, we found the laptop arrangement inefficient: we were limited to a fixed number of individuals that we could test at any one time based on the number of laptops that we could dedicate to the experiment.

3 Web-based Software

We ported the experiment software to a single server so that we could run multiple test sessions asynchronously from different test sites. Portions of the code were rewritten so that an administrator could check up on the subjects' responses and their

[4] The exceptions that we are aware of are Resnik (1997), Levin et al. (2000), and Voss (2002).

[5] BLEU scores range from 0 to 1 (where 1 corresponds to a "perfect" translation), based on n-gram matches with one or more of the reference translations of the source documents. NIST n-gram scores are not normalized within a fixed range.

progress, online while the user sessions were underway. We also introduced a more extensive, online training phase, with a screening procedure that removes subjects from the study if they fail to perform the task correctly prior to the actual evaluation.

The flowchart in Figure 1 for our software, WebLT, traces the sequence of phases that each subject follows in the experiment: introduction, log-in, training, and the actual task. The screen shot in Figure 2 shows the layout of the machine-translated text and category list that a subject sees during the training and actual task phases. During the introduction, the subject reads about the categorization task and procedures to complete the study successfully. Each subject then logs in their user name and is assigned a randomly generated unique ID used to track their data.

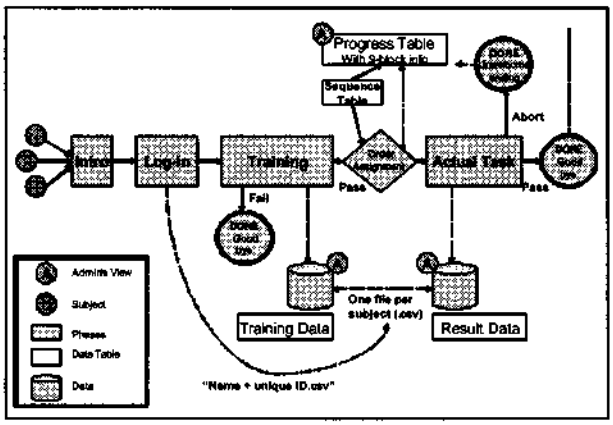

Figure 1 Flow Chart of Phases in WebLT Software

At the next phase called *training,* each subject is introduced to a definition of each category and is presented with sample documents to categorize. As they learn how the software works and categorize each of these documents, they are given feedback online with the correct category answer and an explanation. The subjects are also shown machine-translated texts at this point to familiarize them with text that is not fluent English. Following the initial categorizations, subjects receive feedback identifying phrases from these translated texts that provide the evidence for the texts' categories.

Once the sample documents have been presented, each subject is given a simple screening test that requires correctly categorizing at least five out of six new machine-translated documents. Subjects who do not meet this cutoff are given feedback on their errors and then a second screening test with the same cutoff criteria.

The software permits only subjects who pass one of the screening tests to continue on to the next phase, the actual task. Each subject is assigned the next available sequence number from the *Sequence Table.* This table is a matrix of document code sequences, where each code identifies a test document to be presented to a subject by its category, its replicate id within the category, and the MT engine that generated it. During an experimental session, the administrator is able to examine the *Progress Table* that displays the subjects' sequence numbers, their unique IDs, and a flag that indicates when each user session has been successfully completed. Once the subject successfully finishes the entire experiment, the progress table is updated. If for any reason the subject leaves during the actual task phase, their data collected so far is retained and their session flag in the progress table is marked 'not complete'.

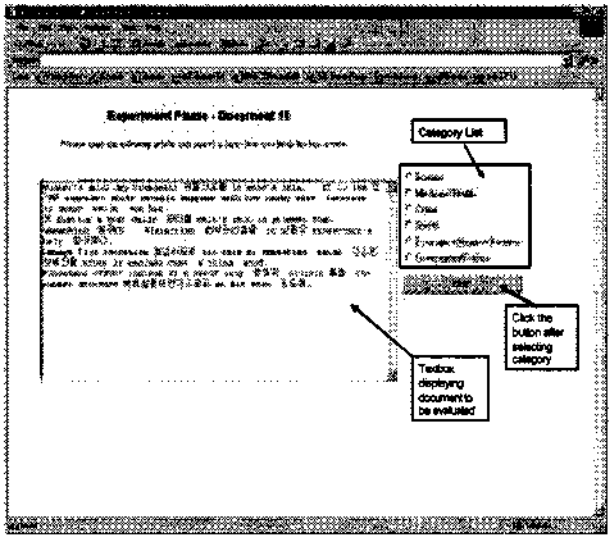

Figure 2 Sample Screen Shot from Experiment

All data collected during the experiment, shown as *Training Data* and *Result Data* files in Figure 1, are stored in comma-separated-value format (.csv). For each subject, there is a separate .csv file that the administrator can see with unique ID, sequence number, sequence of document codes and subject's responses (incorrect or correct match) and time stamps. The software also collects and time-stamps all intermediary responses that subjects may select before they advance to the next document.

4 Experimental Design and Results

For this study, we had three Korean-English MT engines to evaluate. To make sure that all subjects saw multiple documents from each category as translated by each of the three engines, for a total of 18 different source documents, we reduced the original category choices to the following six: science, economics/financial/business, sports, politics/government, crime, and health/medicine. Six or more online Korean news articles were selected for each of six categories from the various Korean sites including www.hankooki.com, and www.joins.com. Each document was then trimmed from the end up to fit the size of the test window, resulting in roughly comparable text passage lengths across documents and categories. The documents were translated by each of the three Korean-to-English MT engines to form the training and test document collections.

All nine participants in the experiment were college freshman who reported that they were 18- and 19-year old science majors. All subjects completed a written questionnaire about their daily use of computers with different operating systems and browsers, their source of news, and the languages other than English that they know or have studied. Five of the nine students had previously studied a foreign language, but not Korean, the source language of test documents. All nine reported using computers two or more hours daily, as well as experience with Windows and the Internet Explorer browser under which our software was running. While all subjects reported following the news, less than half said they read the news over the web. All subjects passed the first screening test of the training phase with a *perfect* six correct out of six category choices. As programmed, the software used this performance to pass all subjects from training on to the actual experiment.

All participants were then asked to categorize each document in one of six given topic categories mentioned above. Each subject was presented with a randomized sequence of these documents, sampled over the 6-topic categories, with three documents per category and six per machine, for a total of 18 total document viewings by each participant. No subject saw the same document translated by more than one system. In addition to meeting the constraints on subjects seeing enough machine-document translations in each category, this design was also chosen to ensure that all machine-document translation combinations were seen by three subjects.

All subjects completed the actual task without complications. The spread of the results were as follows: one subject had a perfect score of 18 correct categorizations on all 18 documents judged, one subject categorized 17 documents of 18 correctly, one subject categorized 16 of 18 correctly, two subjects correctly categorized 15 of 18, and the remaining 4 students each categorized 14 documents correctly out of the 18 they judged.

With each of the nine subjects seeing three documents from each category, there was a maximum of 27 possible correct matches for each category. The category-correct-match responses ranged from 21 on the lower end to 24 on the upper end. The "sports" category was the most frequently judged incorrectly across the systems, while the "government/politics" and the "medicine/health" categories were most frequently judged correctly across systems.

Category	C1	C2	C3	C4	C5	C6	Total
Machine							
System A	8	7	7	9	6	4	41
System B	8	8	8	8	9	9	50
System C	7	8	9	7	7	8	46
Total	23	23	24	24	22	21	137

Table 1 #Correct Responses by System and Category

With each system translating three documents per category and three subjects judging each document translation, there was a maximum of 54 possible correct matches for each machine. These results along with the total correct responses by systems in all six categories are listed in Table 1.

5 Analysis

In this section we test for an MT system effect in our data, progressively refining our interpretation, first with a chi-squared test, then a log-likelihood ratio test (LRT), and finally with an alternate chi-squared test. In the next section we follow up the system effect found here and perform generalized linear model-based (GLM) hypothesis tests for differences among other parameters, such as document-category and subject, to assess their contributions to the experiment's results. This use of GLMs sets the stage for more complex analyses

(A)	INC	COR	TOTAL
System A	13	41	54
System B	4	50	54
System C	8	46	54
TOTAL	25	137	162

(B)	INC	COR	TOTAL
System A	8.333	45.667	54
System B	8.333	45.667	54
System C	8.333	45.667	54
TOTAL	24.999	137.001	162

(C)	INC	COR	TOTAL
System A	2.6138	.477	3.0908
System B	2.2531	.0024	2.6642
System C	.0133	.8905	.0157
TOTAL	4.8802	1.3699	5.7707

Table 2 (A) observed frequencies, (B) expeccted frequencies, (C) chi-squared contribution

that permit the inclusion of a *continuous* variable in the model, such as the non-task-based BLEU n-gram scores, as a predictor of experiment's task-based results.

5.1 Initial Test with Chi-squared Model

We start with the *chi-squared* or χ^2 goodness-of-fit test to determine if there are any statistically significant differences among the MT systems in the experiment. By collapsing responses across subjects and document-categories, we can ask whether the response data show that the probability of correct-responses is independent of the MT system. We test the null hypothesis that success—as measured by the number of correct responses that subjects produced in response to their reading of translated documents in the experiment's task—does not depend on the particular MT system that translated those documents, whether system A, B or C.

The goodness-of-fit test using the χ^2 distribution compares the observed correct and incorrect response-counts (O) to the expected correct and incorrect cell-counts (E), assuming homogeneity of accuracy rates with each MT system for all document-categories and subjects.[6] Using the formula: $\sum_{i,j}(O_{i,j}-E_{i,j})^2/E_{i,j}$ with i rows and j columns, we compute the chi-square statistic as shown in Table 2. For significance at .05 level, the chi-square should be greater than or equal to 5.99.

The resulting chi-squared value of 5.7707 does not point strongly towards rejecting the hypothesis stated above, giving us only weak support for claiming a relationship between MT systems and probability of correct response. However, since the test procedure that we used collapsed the response data into simple correct or incorrect categories without looking at interaction effects of system, subject and category data, it is possible that such effects may have been masked.

5.2 Pair-wise System Comparisons

In some settings, multiple-comparison tests may be more powerful than an overall chi-square for detecting system differences. Given the relatively weak results of the analysis above in determining differences between MT systems, we also tested a second approach to response data, still collapsing across subjects and document-categories, but this time comparing the response data only from the systems pair-wise, i.e., with the correct and incorrect response-counts to documents translated by system A vs. B, by system B vs. C, and A vs. C, in three independent pair-wise tests.

Statistical procedures such as these involving multiple tests, however, allow many opportunities to obtain apparent significance by chance. Therefore, it is desirable to permit only an overall .05 probability for a significant result in the full experiment test. We ran the pair-wise comparison tests at the "Bonferroni-corrected" significance level of .05/3. (Another way to express this correction is that the smallest of the three pair-wise comparison p-values was multiplied by 3 in order to obtain the experiment-wide p-value.)

For this approach, we used the *Log-likelihood Ratio Test* (LRT) (see Agresti, 2003) with the observed frequencies observed in Table 2 part (A). After obtaining the LRT statistic value and associated p-value for each pair-wise comparison, we found only one comparison to be slightly significant, that of system A vs. system B, as shown in Figure 3.

[6] This assumption is open to question, which is why we opt in Section 6 to use a more detailed parametric model like the GLM to test for this homogeneity.

$$2*[13*\log\frac{13}{54}+4*\log\frac{4}{54}+17*\log\frac{17}{108}+41*\log\frac{41}{54}+50*\log\frac{50}{54}+91*\log\frac{91}{108}]$$

$$= 5.908$$

Figure 3 Calculation of hypothesis test for pair-wise equality of MT system accuracy rates (1-degree of freedom p-value: 0.015, used Bonferroni-corrected significance level .05/3=.01667)

5.3 Chi-squared Analysis Revisited

In the pair-wise comparisons of section 5.2, we found that system A and system B could be teased apart, while systems A and C showed no evidence of difference. These results, together with the overall chi-squared result in section 5.1, suggested that system B in particular would yield a higher probability of correct responses than either system A or C. We validated this by running a new contingency-table test for lack of association, with a revised hypothesis that responses to system A and C are identical to each other but different from the responses to system B. Table 3 shows the new counts for this comparison. The new χ^2 value with 1 degree of freedom is 3.9968 which yields a p-value of 0.045 (less than .05), indicating that this distribution is significant. Thus, the probability of correct responses to system B is higher than the probability of correct responses from the other two systems, with statistical significance.

	INC	COR	TOTAL
System B	4	50	54
System A and C	21	87	108
TOTAL	25	137	162

Table 3 #Correct Responses by System

6 Toward Validation of Approach

Given the task-based results in ranking of the MT engines, we have begun to investigate whether a *non-task-based* metric can independently validate the ranking. For example, we would like to know if it is possible to find a correlation of such a metric with task-based results, using statistical regression and GLMs.

6.1 Generalized Linear Models (GLMs)

In a simple linear regression model (SLR), the relationship between two variables is modeled by fitting the collected data to a linear equation. With one variable as explanatory variable and the other as dependent variable, the objective of SLR is to determine whether a model can be found to fit the data and predict experimental behavior. The generalized linear model (GLM) is a generalization of the linear regression model that can be used to fit linear or nonlinear effects of predictor variables, whether categorical or continuous (Agresti, 2003).

The basic GLM model is represented as the following *linear predictor function* that models the relationship between the response variables Y_i and explanatory variables x_{ij} :

$$\eta_i = \sum_j \beta_j x_{ij}$$

with assumptions that (i) the Y_i are independent responses, (ii) Y_i is a random variable with an exponential family distribution, and (iii) the mean $E(Y_i) = u_i$ satisfying:

$$g(u_i) = \sum_j \beta_j x_{ij} = \eta_i$$

6.2 Evaluation of Task Results with GLMs

For our data, the GLM components consisted of three *explanatory variables* (subjects, categories of documents, and MT systems) and two *response variables* (number of correctly matched responses and number of incorrectly matched responses). We used S-PLUS, a statistical software package, for the computation of this analysis (Venables and Ripley 1999). First, with a data frame matrix of rows for system *(sys),* category *(catg),* number of correct matches within category *(succ),* and number of incorrect matches within category *(fail),* the data was assessed for the category contribution to correct matching and fit with a logistic regression model with *catg* and *sys* as categorical predictor variables.

The *catg* coefficients in the model were extremely small relative to the other coefficients in the model. Thus this test showed no category effect. We note that the estimated regression coefficient for system B had the most outstanding value and, when contrasted with that of system A their difference was also quite significant.[7]

We also considered another explanatory variable, the subjects. Using model 2, a revision of the first model was run with a data frame matrix of rows

[7] This is valid only because system A and B coefficient estimators were found to be approximately independent.

for system *(sys),* subject *(pers),* number of correct matches within subject *(succ),* and number of incorrect matches within subject *(fail).* To assess the subject contribution to correct matching, we fit another logistic regression model. This time the results suggested that two variables were significant, system B and subject 2 *(pers2)* due to their exceptional regression coefficients.

6.3 Evaluation of Task Data with BLEU Scores

As a first step in looking for a non-task-based metric to independently validate a task-based approach to MT evaluation, we ran the BLEU metric with three reference translations on the documents from only one category of the task dataset, business. Figure 4 displays the BLEU scores on this category, with n-gram lengths of 1 through 4. The n-gram metric produced an MT system ranking that is fully compatible with the one derived in the task-based experiment, where system B generated the translations that subjects were most likely to categorize correctly. Figure 4 shows that the BLEU-based ranking remains consistent with the task-based ranking at each of the different n-gram lengths tested.

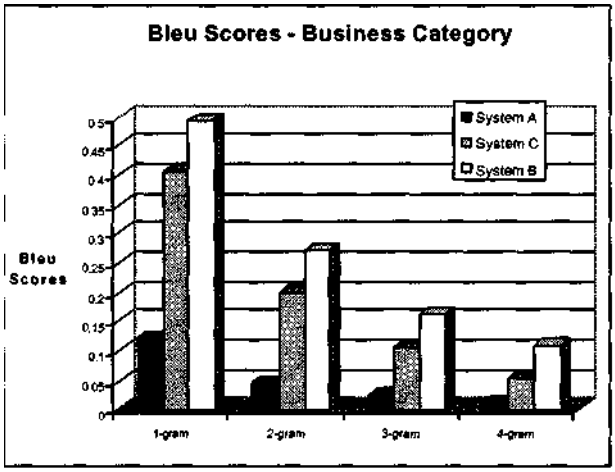

Figure 4 Bleu Scores on Task Dataset

On the strength of these results, we then built the full set of three reference translations for all documents in the other five categories and ran the BLEU 4-gram metric on these categories as well. As Figure 5 shows, the BLEU scores by category for the rest of the task dataset consistently yielded the same results found within the business category: in short, the BLEU-based ranking of the systems was fully compatible with our task-based ranking results.

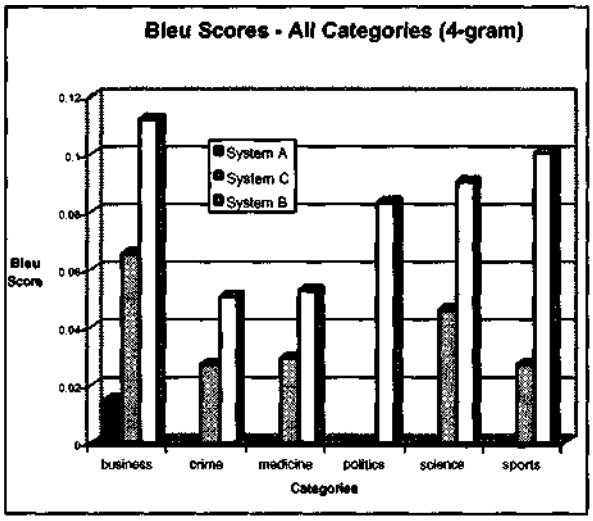

Figure 5 Bleu Scores on Task Dataset

7 Conclusion & Future Work

Our work in developing a practical, language-independent method of evaluating MT engines to assess their realistic use in actual tasks of interest to our users, led us to focus on three distinct, but interrelated aspects of evaluation. First, we constructed *webLT,* an interactive, server-based software program that enables the evaluators to administer the experiment while the users who participate in the experiment perform their actual task. The software tracks and records the users' online actions, timing data, and task decisions. Second, we reviewed various designs and selected a straightforward factorial experimental design with replicate observations for its simple manageability, enabling us to build and readily assign a tractable-sized set of translated test documents in different categories to a relatively small number of users. The assignments were put in a sequence table within *webLT* and can be readily re-used or modified in future experiments. Third, we found that with available statistical software tools, we could move beyond chi-squared tests, as needed, to work with generalized linear models (GLMs) to permit finer-grained analyses of our data.

This combination of software, experimental design, and statistical models for data analyses was

used to rapidly and successfully assess three Korean-English MT systems in terms of users' performance accuracy on a categorization task. The experimental results showed that documents translated by System B were categorized correctly with a statistically significant, higher probability than those translated by Systems A and C. The experiment proper took less than three hours total and could have been run in less time had more people been available simultaneously.

Although the task dataset is small compared to NIST MT evaluation test sets, when scored with three reference translations using the BLEU n-gram metric, we found that the BLEU-based ranking of the systems was fully compatible with category and at different n-gram lengths, with our task-based ranking results. Future work will involve testing user performance on linguistically more complex tasks, such as identifying named entities and event types, from translated texts. Further work will also include a more extended set of analyses with Bleu scores for the test documents, where the scores will be treated as explanatory variables within the GLMs, to assess their strength as independent predictors of specific task results. The question will be whether task-based results can be predicted by Bleu scores (or vice versa) within a statistical model, in order to provide independent validation for both.

Acknowledgements

The development and implementation of the WebLT software was a joint effort with ArtisTech, Inc. in ARL's Collaborative Technology Alliance. Dr. Eric V. Slud of the Mathematics Department at the University of Maryland provided experimental design recommendations and extensive feedback on the statistical analyses of this paper. We thank students in U. of Maryland's CMPS299A class for participating in the study and one reviewer for a detailed set of valuable questions on the analyses.

References

Agresti, A. (2003) Categorical Data Analysis. John Wiley & Sons, Inc., NJ.

Church, K.W. and E.H. Hovy. (1993) "Good applications for crummy machine translation" *Machine Translation,* 8, pp. 239-258.

Doddington, G. (2002) "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics." In *Proceedings of HLT 2002,* Human Language Technology Conference, San Diego, CA.

Hovy, E.H., E. Filatova, M. King, and B. Maegaard (2001) "The ISLE Classification of Machine Translation Evaluations." Report at http://www.isi.edu/natural-language/mteval

Hovy, E., M. King, B. Maegaard, and M. Palmer (2003) "FEMTI, the Framework for Machine Translation Evaluation in ISLE," document at www.issco.unige.ch/projects/isle/taxonomy3/

Levin, L., B. Bartlog, A. Llitjos, D. Gates, A. Lavie, D. Wallace, T. Watanabe and M. Woszczyna (2000) "Lessons learned from a task-based evaluation of speech-to-speech machine translation" *Language Resources and Evaluation Conference* (LREC), Athens, Greece.

Melamed, I.D., R. Green, and J.T. Turian (2003) "Precision and recall of machine translation" *Proceedings of Human Language Technology and North American ACL,* Edmonton , Canada.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2001) "Bleu: a method for automatic evaluation of machine translation", *IBM Research Report* RC22176, Yorktown Heights, NY.

Pastra, K. and H. Saiggon (2002) "Coloring summaries Bleu" In *Proceedings of the EACL 2003 Workshop,* "Evaluation initiatives in NLP: are evaluation methods, metrics and resources reusable?" Budapest, Hungary.

Resnik, P. (1997) "Evaluating Multilingual Gisting of Web Pages", *Proceedings of AAAI Symposium on Natural Language Processing for the World Wide Web,* Stanford, CA.

Taylor, K. and J. White (1998) "Predicting what MT is Good For: User Judgements and Task Performance." *Proceedings of the Association for Machine Translation in the Americas* (AMTA-98), pages 364-373, Lansdowne, PA.

Venables, W.N. and B.D. Ripley (1999) *Modern Applied Statistics S-PLUS.* Springer-Verlag NY.

Voss, C. (2002) "MT evaluation: measures of effectiveness in document exploitation." *Presentations of the DARPA TIDES PI Meeting,*
Santa Monica, CA.

Zajic, D., B. Dorr, and R. Schwartz (2002) "Automatic headline generation for newspaper stories." *Proceedings of the ACL 2002 Workshop on Automatic Summarization,* Philadelphia, PA.