Toward a Predictive Statistical Model of Task-based Performance Using Automatic MT Evaluation Metrics

Abstract

Can automatic intrinsic metrics predict when an MT engine is "good enough" to support people performing sense-making tasks, such as the extraction of *who, when,* and *where* information from translated documents? This paper reports on the challenges and results of a large-scale, task-based experiment to evaluate three Arabic-English MT engines with subjects' response rates (extrinsic metric), and on the new application of statistical model specification, parameter estimation, and adequacy assessment, to *predict* these response rates from descriptive experiment variables and a novel, automatic intrinsic metric derived from BLEU, to control for translation difficulty of test documents. The best logistic regression model to date provides a new method that is both mathematically rigorous and independent of subjective human judgments, for pursuing the goal of ultimately replacing slow, expensive task-based evaluations with faster, less costly automatic intrinsic evaluations.

1 Introduction

Automated ngram-based metrics have been widely accepted for intrinsic MT evaluation among MT researchers because the metrics correlate well with human judgments of translation quality and they yield rapid feedback during development of MT engines. As a result, MT users now want to know whether automated metrics will help them to calibrate MT engines for the the range of real-world tasks that they must perform. Before calibration is feasible however, the field of MT evaluation needs to develop (i) *extrinsic metrics* for users' task-based performance and (ii) statistical models to predict such metrics from *automatic intrinsic metrics.*

This paper addresses both of these research areas, reporting on (i) the results of one large-scale, task-based experiment with analyses of human performance data to compute response rates as extrinsic metrics and (ii) the statistical models that have been tested for predicting the experiment's response rates from *BLEUavg*, a novel, automatic intrinsic metric derived from BLEU, to control for translation difficulty of test documents.

The experiment—one of a set of several that were designed to test progressively more complex tasks that real users perform, as originally hypothesized in a task hierarchy by Taylor and White (1998)—required subjects to extract phrases with *who, when,* and *where* information from machine-translated newswire texts. The research focus was to establish a task-based method and extrinsic metrics for determining how effectively distinct types of Arabic-English MT engines can support people on the phrase extraction task.

We report on two sets of results, those from the extraction experiment in establishing response rates as extrinsic metrics, and those from new research in constructing statistical models to predict the task-based rates from the experiment data. While two of the three MT engines under evaluation in the experiment yielded statistically indistinguishable "hit (accuracy) rates" from subjects for phrase extraction from MT output, there were statistically significant interactions between the type of wh-phrase being extracted and MT engine, as determined by chi-square tests. This results provides support for including this extraction task as a cut-point in the hierarchy, with which weaker MT engines can be distinguished from stronger ones. The second set of results reported in this paper both introduces the process of constructing and testing statistical models for cross-classified data and presents the sequence of models tested on the experiment data. To date, the best predictive model is a logistic regression function with eight parameters. While this model does not yet achieve statistical adequacy (where the discrepancy between observed and predicted values is still too systematic to be entirely due to chance), we have reached a point where the model-fitting is at a favorable, even if not quite final, stage.

2 Background

Two strands of ongoing research are related to our work. First, task-based evaluation of MT engines have been conducted for many years, but the extent of the statistical analyses and the experimental designs has been limited by the scale of the studies (Taylor and White 1998, Doyon et al. 1999). Recent projects have increased subject pool size, indicating a new interest on the part of government agencies to ensure that resulting datasets are large enough to support statistical significance testing (Jones et al. 2005, Voss et al. 2004). Given that task-based evaluation of software tools generally is not uncommon outside of MT as part of the broad category of user-centered evaluation, where the goals are to involve the users early on in the development process to gather the requirements of their tasks and then at later stages to bring them back to evaluate the utility and usability of the tools (Damianos 2001), it is curious that not more task-based MT evaluation has occurred since Church and Hovy (1993) wrote that there were "good applications for crummy MT."

The other strand is the development of new automated intrinsic metrics (Papineni et al. 2002, Doddington 2002, Melamed et al. 2003, Lavie et al. 2004, Lin and Och 2004) spurred by MT developers themselves. Until now, the validation of these automated evaluation metrics has been in terms of human subjective judgments of MT output. Little has been done in terms of comparing automated metrics and task-based performance.[1] Also this validation work has been done with correlation analyses. Other techniques in mathematical data analysis, such as working with generalized linear models (Dobson 1990) that are more powerful and sensitive, has yet to be explored.

3 Overview of Approach

In this section we briefly describe the extraction task of our experiment with the analyses performed on the collected subject response data, the data structure with the variables stored for statistical modeling, and a new metric called BLEUavg, introduced to capture document translation difficulty.

3.1 Extraction Task

Of the various text-handling tasks that Taylor and White (1998) ranked for level of required translation quality, extraction is the most specified linguistically, with simpler syntactic structures at the shallow end and more complex relations at the deep end, as reproduced from their paper here in Table 1. In examining the outputs of the three MT engines to evaluation, we were able to identify noun phrases and prepositional phrases, but we were not as confident of the roles of different phrases in the sentences' argument structure. As a result, we defined our experiment task at a slightly more general level than Taylor and White's "shallow" extraction. Rather than just extracting the named entities that are proper names, we broadened the categories to essential elements of information and tested for how well subjects could identify people, places, and times in machine-translated documents. Their task was to highlight all of the "Who", "When", or "Where" expressions as requested in the displayed documents.[2] Our hypothesis was that people would perform best with those MT engines that both preserve these phrase types as contiguous units in translation and generate the correct English word order within the units, if they need to "read for phrases" in MT output; i.e., go beyond word-scanning and identify sequences of words that contain `who`, `when`, or `where` content.

3.1.1 Identifying the Correct Answers

In order to identify the correct answers in the machine-translated documents, four native Arabic speakers, all bilingual in Arabic and English, translated the training and test documents into English to create the four reference translations for the corpus. The identification of Who, When, and Where ground truth (GT) elements in the Arabic texts was established by one native Arabic translator. These items were then discussed with a trained linguist who, working with the English human reference translations, marked up these documents for their parallel reference translation (RT) phrases. The tracking of RT items in correspondence with their GT items was recorded in a spreadsheet as the translator and linguist working together reviewed each source document. RT mark totals (`RTMTot`) were recorded for each document. (At the statistical analysis stage, RTMTot counts are treated as design constants of the experiment.)

Figure 1 displays a short Arabic phrase with the GT item underlined and the reference translation with the corresponding RT item underline beneath it. After the GT-RT items were established, the following procedure for identifying the corresponding "omniscient truth" (OT) elements in the MT outputs occurred. Given a listing of the RT elements by document in order of appearance within each sentence of the document, the annotators searched within the same sentence of the MT output text for the OT item that best approximated the RT element. The OT "chunks" were selected semantically by the annotators, so that even when they found incorrect English syntax or incomplete translations only roughly corresponding to the RT element, they could identify an OT item.[3]

[1] However, Dorr et al. have investigated this in the field of summarization evaluation.

[2] On any document, subjects only extracted one type of wh-item.

[3] The best phrase translations were categorized as 'A's, the

Type of Extraction	Description
Deep	Event Identification (scenarios): the ability to identify an incident type and report all pertinent information
Intermediate	Relationship identification: (member-of, associate-of, phone-number-for)
Shallow	Named entity recognition: (isolation of names of people, places, organizations, dates, locations)

Table 1: Multiple Levels of Extraction

The set of OT items for a document at times varied with the MT engine that generated the output text in the document. This can be seen in the example in Figure 1 where the underlined subject of the verb is translated and effectively "lost" by MT3, while it is transliterated by both MT1 and MT2, and then separated across the verb or "split" in MT1.

3.1.2 Measuring Subjects' Task Accuracy

The subjects' responses were collected via a browser interface as they clicked on the displayed translated text to select words or phrases as their answers. The responses were then encoded directly inside copies of the text with begin- and end-tags. These tag-marked responses were then matched against the omniscient truth (OT) answers tagged the same way in the previously prepared answer texts. The subjects' responses were classed as "Hits" (where the subjects tagged phrase fully matched an OT answer), as "Misses" (where the subject missed or failed to tag any portion of the phrase that was an OT answer), and "False alarms" (where the subjects tagged phrase text was not included in OT answer set).[4]

3.2 Data Structure of from Task-based Experiment

In the information extraction task, documents are grouped into one of 3 WH-types (Who, Where, When). Each subject extracts the WH-type from each document according to the particular group for which it is associated. For example, subjects search for all relevant *who* phrases in a document from the **Who** category. Each group of documents include six **replicates** or copies. Thus, an individual document is indexed by category *WH*

GT: كتبت ريم الميح: في قصر بيان....
RT: Reem Meeh wrote:
yesterday at Bayan Palace...
MT1: Reem wrote 'Lmeeaa : [S]
in a statement derelict , ...
MT2: I wrote Rim almyai [الميح] : [A]
in the short statement, ...
MT3: clerks move flowing : [Z]
in castle demonstration/statement? .

Figure 1: Sample Parallel Phrases in Parallel Texts: Who ground truth (GT) phrase in Arabic source text, Who reference truth (RT) phrase in reference translation, Who omniscient truth (OT) phrases

and replicate number *Rep*. Each distinct document is translated by each of the three MT systems, numbered 1, 2, or 3, and each translated document, indexed (WH, Rep, MT), is viewed the same number of times by individual human Subjects. The total number of **Cases**, or translated-document by Subject records, is equal to the product of the number of Categories, the number of Rep's, the number (3) of MT systems, and the number of subjects who see each physical translated document. This particular experiment had been designed for 60 Subjects, but was conducted with 59 because one candidate was a no show. As the experiment was actually conducted, 57 subjects saw 18 documents, and 2 saw 17.

The experiment data falls into two major categories as defined in statistical modeling: *Response* variables—Hits, Misses, False Alarms—which are the variables of interest that are being predicted in the model and *Explanatory* variables—WH, MT, automated metrics, subjects—which are important in making the prediction about the response. The rate of the *Hits* response is being modelled in this paper. The explanatory variables in our model consist of what are called *dummy* variables for subjects, WH-categories, and MT systems, together with recoded BLEU variables and pairwise *interactions* or products of these variables, for each translated case. Statisticians call the items that are used for categorical

phrases that were only partially correct were 'B's, the phrases whose elements were split apart by words that did not belong in the phrase were 'S's, and the phrases that were dropped out of the translation were 'Z's. Each of the annotators' OT items and their ABSZ codes were also placed in the spreadsheet, lined up with the corresponding GT and RT item. Inter-annotator scores (Carletta 1996, Di Eugenio and Glass 2004) were calculated as we refined the annotation process. The scores for the 3 MT engines on each of the wh-type documents were all within the 0.6 to 0.8 "good agreement" range. Three of the nine Kappa scores for within who/where/when types were above 0.8 in the "very good" range.

[4]While the terms "Hit", "Miss" or "False Alarm" are taken from signal detection theory (SDT), we note that the usage in Task 2 is technically distinct from the usage in SDT. When subjects marked part of an OT answer but did not include all the content words, their response was coded "Partial," not a hit.

variables "dummy variables" or "binary indicators" — what might be called a "boolean variable" by computer scientists—which have value 1 if true for case and 0 otherwise. For categorical variables, the *main effects* are the binary dummy-indicators and the *interactions* are pairwise products involving dummies with each other and with BLEUavg.

3.3 Automated Metric Choice

We define a composite documentwise BLEU4g score called BLEUavg by taking ordinary arithmetic averages across MT systems of the ratios of BLEU4g scores divided by the within-MT average over documents. These composite scores showed the most promise among a number of such re-coded variables we considered. We use them as predictor variables in the statistical model for event rates, which we describe next. As mentioned earlier, we are aware of other metrics that have been proposed and we ultimately plan to consider each in our modelling as well.

The most widely used automatic evaluation metric is the BLEU 4-gram metric (BLEU4g). Because of familiarity with use and for results comparable to other evaluations, we chose to use this metric in our analysis. In the course of model building, BLEUavg was introduced to create a metric that will be more successful in drawing conclusions about documents for difficulty or *translatability* to be used in predictive modeling. Translatability refers to attributes of document quality and will be a very interesting predictor (in interaction with other predictors) for translation success. This concept of quality determines whether a document is easy or hard on a given scale of difficulty. Thus, a system could be evaluated based on how well it translated documents of a specific level and furthermore, based on how well certain tasks are performed in relation to the documents' degree of difficulty. This exploits the notion of quality vs. usefulness.

4 Statistical Methods

Statistical modelling can be used to summarize experimental data. One must take into account the following information for model fitting: (Dobson 1990):

- Model specification – the mathematical formula chosen to relate the expected response (hit-rate) to the explanatory variables, and the probability distribution of the response.
- Parameter estimation – the method of estimation (maximum likelihood), including numerical algorithm and output summary statistics, which are all usually standard in statistical software packages for widely used models like generalized linear models.
- Model adequacy – the statistics and diagnostic exhibits to be used in measuring the fit of each model to the data.
- Inference – confirmatory conclusions and interpretations that can be drawn from model parameters.

One effective statistical modelling framework for relating automatic evaluation metrics to human task performance is regression, specifically logistic regression.

4.1 Logistic Regression for Predictive Modeling

The generalized linear model (Agresti 2002), or GLM, is a generalization of the ordinary linear regression model. As in linear regression, the expected response variable $\mu_i = E(Y_i)$ for the i'th observation or case is specified as a known function of a linear combination $\eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ of the explanatory variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ in the i'th case, where the coefficients α (called *intercept*) and β are unknown parameters, common to all observations and estimated from the data. While $\mu_i = \eta_i$ in ordinary linear regression, a GLM requires further specification of a known *link function* g to capture the relationship between $E(Y_i)$ and $\mathbf{x}_i$, in the form

$$g(\mu_i) = \eta_i$$

The modeler chooses the function g as well as the form of the probability distribution of the response Y_i (conditionally given η_i), from a so-called *exponential distributional family* (Agresti 2002)

In our setting, $Y_i = \texttt{Hits}_i/\texttt{RTMtot}_i$ is modelled as though $\texttt{RTMtot}_i$, $\mathbf{x}_i$ were fixed; the number $\texttt{Hits}_i$ of hits is treated as the number of successes in $\texttt{RTMtot}_i$ independent coin-tosses, each with success-probability given by $\mu_i = E(Y_i)$ satisfying $\eta_i = g(\mu_i) = \log(\mu_i/(1-\mu_i))$. This choice, associated with the particular GLM called the *Logistic Regression Model*, is called the *logit* link-function along with the *Binomial* distributional family, and implies the relationship

$$E(Y_i) = \frac{\exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})} \quad (1)$$

The goal of the logistic regression model, like all GLM's, is to express the predicted or expected behavior of the response variable Y_i through an explicit parameterization (1) in terms of the explanatory variables. The parameters are readily estimated from data using the method of maximum likelihood, standardly implemented in the **R** statistical package (and almost all others).

4.2 Predictive Modeling for MT Evaluation

In this setting, a "predictive model" is a parametric formula for the event-rate (such as Hit-rate) in terms of a specified combination of document characteristics such

as what MT system translated the document, the label WH describing whether the document was classified according to *who*, *when* or *where* items, and the document difficulty index BLEUavg recoded by averaging BLEU over MT systems. The unknown (α, β) parameters are estimated by maximum likelihood (ML). GLM model-building techniques as described in the following paragraphs are useful in determining just which explanatory variables and interactions enter into the prediction of event occurrence probabilities. The resulting logistic regression model provides a finer-grained analysis of effects in the information extraction task than would a simple display of event-rates.

We must determine which variables enter as significant predictors into a GLM describing the performance of event rates cross-classified by MT, WH, and various automated evaluation metrics. The number of terms we enter into the model are often changed or updated after looking at the data during model fitting. The idea is to include terms in the model which are highly significant as opposed to barely significant, which enables us to avoid fitting noise after examining many possible terms to include. An iterative approach to model building, testing, and fitting is being explored in this research. We utilize a model-building strategy that will eventually give new information about the relative predictive strengths of the different automated metrics (and possible new variants of them).

For the purpose of MT evaluation, statistical models have several important uses. The first is predictive: after fitting and validation of the model, the investigator can use the model to determine a best-guess or likely range of event-rates that would arise with a new combination of documents, MT engines, and human subjects on the same task. A related use of the model is to identify which combinations of document, engine, and subject features will be particularly associated with high task-based event-rates and which with low ones. Even when venturing into a new body of previously untested documents of similar type, the model can be used to approximate the likely event-rates on similar tasks, a very helpful element in designing new data-collections. For example, the sample size in a new data collection, which would be needed to discriminate (with specified probability) at specified precision between the overall performance of two MT engines, can be predicted using the fitted parameter values of such a model. Users of MT engines could apply the form, although not the specific parameter values, of such a model in specifying a similar model to fit on pilot data from a new data-collection for example on a new set of documents and then use *that* fitted model to predict outcomes from a fuller data-collection.

4.3 Model Evaluation

The validity or 'adequacy' of a predictive statistical model, with respect to a dataset, is essentially the property that the observed data deviate from predicted or expected values by no more than would occur by chance some large fraction of the time (say 95%) if the event-occurrence data were actually generated from the given set of predictor variables via the postulated model. There exists an established body of statistical theory and "goodness of fit tests" (Agresti 2002, Ency. of Statistics 3rd Volume) to assess the deviations between observations and predictions from fitted models, taking into account that the fitting of parameters was (or was not) done on the same dataset being used to test adequacy. Statistical models passing such tests of adequacy are the gold standard for applied statistical investigations, because most theoretically based statistical statements (for example, about the uncertainty in an estimated quantity) depend on using a 'correct' model in this sense. Chi-squared (χ^2) tests can be used to measure goodness of fit of models to the data. These tests assess fitted models by measuring discrepancies between observed and expected event counts in a cross tabulation. Alternative logistic regression model fitting methods and other generalized likelihood ratio tests of fit may also be applicable.

5 Stepwise Model Selection

The sequence of models implemented is described in this section. All models described here are in relation to *Hit* event rates. Deviances and/or standardized coefficients determine whether new terms should be added at each step during model building. Stepwise model-building is a standard statistical method and can be implemented in most statistical packages (Agresti 2002, **R** Development Core Team 2004).

5.1 The Simplest Models

Since the idea is to describe the probability of "success" on the WH-extraction task as a function of MT system and/or BLEU scores, as an initial starting value, we attempt to fit the simplest model (Eqns 2 and 3) with only MT and BLEU-style metric main effects.

Model MT only: Model 1A

$$log(\frac{pHits}{1-pHits}) = \alpha + \beta_1 I_{[MT=2]} + \beta_2 I_{[MT=3]} \quad (2)$$

Model BLEUavg only: Model 1B

$$log(\frac{pHits}{1-pHits}) = \alpha + \beta_1 BLEUavg \quad (3)$$

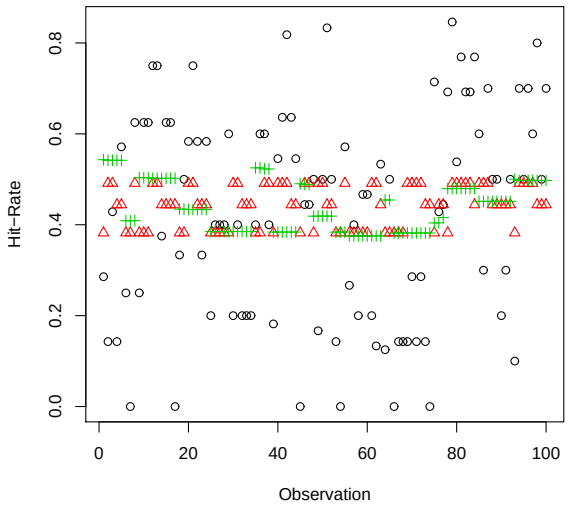

Figure 2: Observed vs. Predicted Hit Rates for MT and BLEUavg only Models for 100 random values. Circle - Observed Rate, Triangle - Model 1A Predicted Rates, Plus - Model 1B Predicted Rates

A few points must be addressed to clarify the models and their notation. In equation 2, two MT system variables, $I_{[MT=2]}$ and $I_{[MT=3]}$, are represented. However, there are 3 MT systems under investigation in this study. These are the dummy indicators used for categorical variables as described in Section 3.2. The **R** software produces $n-1$ linear predictors for each variable, where n is the number of distinct values of the variable. The predictor in the model is interpreted as the increment in the linear score due to a given categorical-variable value versus a baseline value. In our example (2), the term $I_{[MT=2]}$ results in an increment to $\log(pHits/(1-pHits))$ of β_1 over the baseline score for MT=1.

All estimated coefficients in both models are found to be statistically significant, however, neither of the simplistic, one-parameter models work well in predicting the proportions of hit rates. Figure 2 confirms the lack of fit of the two models for 100 randomly selected cases.

5.2 Higher Dimensional Models

The previous models lead us to believe that other variables in our data set must be useful in accounting for the explanation of the Hit rate. Several combinations of the main effect variables and their interactions were tested in different models. At first, it appeared that the most richly parameterized models fit the data best and come closest to the observed experiment hit-rates. However, further examination revealed large standard errors for the estimated coefficients of interaction terms in these models. The interaction terms were highly correlated (multicollinear) with the corresponding main-effect terms in the regression equation, making it impossible to disentangle the relative importance of main- and interaction- effects.

Keeping this in mind, models that support a significant reduction in deviance, have coefficients that are significant, and did well in an external non-model based comparison (i.e. χ^2 of observed vs. expected value) were retained. Results of the best models fitting these criteria include the variables MT, WH, BLEUavg , as well as interactions between WH and BLEUavg, as described below in Models 2A and 2B.

5.3 Results

Thus far, the successive stages of model fitting have been:

- Model 1A: MT only
- Model 1B: BLEUavg only
- Model 2A: MT + WH+ BLEUavg
- Model 2B: MT + WH+ BLEUavg + BLEUavg*WH

Model 1A has 3 parameters (intercept and 2 binary indicators, as shown in Eqn 2), Model 1B has 2 (see Eqn 3), Model 2A has 6 (intercept, 2 indicators for MT, 2 for WH, and 1 for BLEUavg) and Model 2B has 8 (intercept, 2 indicators for MT, 2 for WH, 1 for BLEUavg, plus 2 for the BLEUavg*WH interaction). Table 2 displays the observed counts (found in the $\widehat{\text{H}}$its column) vs. predicted or expected counts (found in columns $\widehat{\text{M}}$od1A-2B)) tabulated across WH and MT. Further, the counts of success are split by a re-code ,called HiBLEUavg(represented HIBavg in table), of the BLEUavg to denote when cases falling into these bins are above (TRUE) or below (FALSE) their overall median values. To check whether this interaction is quantitatively greater than what might occur by chance, we performed a chi-square test for the Hit event-rate. The statistic was obtained in the standard way by summing $(Observed - Expected)^2/Expected$ over all cells of the 18 X 2 table consisting of the 18 WH x MT x HiBLEUavg cells tallying 'non-hitrates' (RTMTot-Hits). The chi-square constructed in this way for Model 2B is compared to the chi-square percentage point with 10 degrees of freedom, calculated by the formula $\widehat{\text{n}}$umber of cells - number of model parameters (in this case, 18-8). These goodness-of-fit statistics show that Model 2B comes closes to adequately representing the proportions of hits in the data, with a chi-square value of approximately 122.17. Even though this value is extreme for a chi-square of the given degree of freedom and shows Model 2B would not be a final model, it is still favorable considering the other model fitting stages. The previous chi-square values for models 1A, 2B, and 2A are

416.00, 346.40 ,and 176.07 on 15, 16 , and 12 degrees of freedom respectively. Chi-square would have to be extremely smaller for the models to be judged 'statistically adequate' which may not be attainable without introducing other structure such as the between-subject variability into the model through random effects.

Figure 3 displays graphically the fit between the predicted values of the 'best' and most parsimonious model, Model 2B, and the observed data. This model is very close to capturing the hit-rate event counts, as seen by comparing its predicted counts versus the observed counts by (WH, MT, and HiBLEUavg) category.

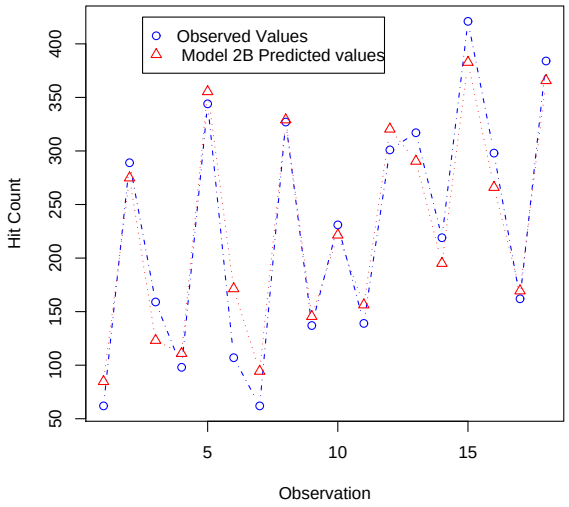

Figure 3: Observed vs. Predicted Hit Counts–WH by MT

It should be noted that a valid model might not be fully predictive in the sense that there might be sources of random variation from human subject-to-subject variability, which could not be modelled specifically. We tried treating each subject as its own individual variable in a model but this model is highly over-parameterized because there are 59 subjects. We also examined several possible groupings based on experimental information obtained from the subjects during the study, like job title, years experience, or experience with MT systems, but none of these variables were closely correlated with subject-specific hit rates. Nevertheless, it would be worthwhile to consider a predictor for subjects in the final model. In this case, such a model might describe the resulting subject-specific increments to event-rate in terms of a random variable with a specified distribution for subject variations, such as a normally distributed variate with mean 0 and variance fitted as a model parameter. Such models are called *random- or mixed-effect models* and will be explored in future model-building.

6 Discussion

After attempts to fit logistic regression models with various combinations of predictor variables, the 'best' model found for the occurrence of Hits turned out to be:

$$log(\frac{pHits}{1-pHits}) = -.809 + .465 * I_{[MT=2]} + .266 * I_{[MT=3]} + .145 * BLEUavg + .035 * I_{[WH=Where]} - .902 * I_{[WH=Who]} + .313 * BLEUavg * I_{[WH=Where]} + 1.07 * BLEUavg * I_{[WH=Who]} \quad (4)$$

Recall that MT=1 and WH=when respectively served as baseline categories for the MT and WH variables. It turned out that the significant contrasts were between both MT systems MT2 and MT3 with MT1 and both Wh-types Where and Who with When. Thus, we included these MT and WH predictors in the final model. In addition, the model predictor BLEUavg proved useful as a predictor for hit-rate and can be interpreted as a measure of the degree of difficulty of a translated document. A logical conclusion of this final model is that the contrast between MT2 and MT3 should be distinguished with respect to success on the information extraction task, as well as the WH-types Where and Who.

7 Conclusion

In this paper we have presented our framework for task based machine translation (MT) evaluation and predictive modeling of task responses. An effective model for predicting *aggregated* hit rates has been developed, even as the *individual subject-marked translated document* hit rates are intrinsically *not* predictable from these models. The comparisons between models show that only models (i) that include main effects for MT, WH, BLEUavg, and those models (ii) that explicitly have a way for hit-rates to vary differently with BLEUavg within different WH categories, show success in reproducing overall hit-rates within the 18 categories defined in Table 2.

The most successful model found does not yet achieve statistical adequacy but still promises to be very useful. We would like to include additional automated metrics (or variants) to create predictor variables introduced for document difficulty, as well as build models to predict the other event rates (i.e., Misses and False Alarms). New insights will develop about metrics, using goodness of fit criteria in models like those developed here, as measurements of predictive task success. Ultimately, we hope to extend MT evaluation methodology to create new metrics specially relevant to task-based comparisons so users can tie the intrinsic automated metrics to the extrinsic metrics for task they perform.

Table 2: Observed vs Predicted Hits totalled over WH x MT x HiBLEUavg

						Models			
Obs	WH	MT	HiBavg	Hits	RTMTot	Mod1A	Mod1B	Mod2A	Mod2B
1	WHEN	1	FALSE	62	260	99.34	103.92	70.58	84.74
2	WHERE	1	FALSE	289	767	293.05	295.84	271.76	274.85
3	WHO	1	FALSE	159	453	173.08	182.04	154.12	123.18
4	WHEN	2	FALSE	98	255	125.25	102.02	94.55	110.97
5	WHERE	2	FALSE	344	754	370.36	291.28	350.51	355.38
6	WHO	2	FALSE	107	460	225.95	184.93	206.45	171.41
7	WHEN	3	FALSE	62	244	108.32	97.40	79.51	94.34
8	WHERE	3	FALSE	327	780	346.27	301.14	325.26	329.05
9	WHO	3	FALSE	137	444	197.11	178.51	178.37	145.51
10	WHEN	1	TRUE	231	621	237.27	304.41	237.59	221.51
11	WHERE	1	TRUE	139	340	129.91	160.96	160.70	156.30
12	WHO	1	TRUE	301	650	248.35	310.62	286.27	320.42
13	WHEN	2	TRUE	317	620	304.54	303.30	304.74	290.44
14	WHERE	2	TRUE	219	340	167.01	160.50	198.52	195.04
15	WHO	2	TRUE	421	637	312.89	303.27	351.23	382.77
16	WHEN	3	TRUE	298	635	281.90	309.90	281.04	266.01
17	WHERE	3	TRUE	162	323	143.39	152.48	173.26	169.38
18	WHO	3	TRUE	384	660	293.00	314.48	332.57	365.71

References

Alan Agresti. 2002. Categorical Data Analysis. John & Wiley Sons, Inc., NJ.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, June.

K. Church and E. Hovy. 1993. Good Applications for Crummy Machine Translation. *Machine Translation*.

Laurie Damianos. 2001. User-Centered Evaluation. Briefing, MITRE Corporation. [http://www.mitre.org/work /tech_papers/tech_papers_01/damianos_evaluation/ damianos_evaluation.pdf].

Barbara Di Eugenio and Michael Glass. 2004. Squibs and Discussions - The Kappa Statistic: A Second Look. *Computational Linguistics*, pages 95–101.

Annette J. Dobson. 1990. An Introduction to Generalized Linear Models, Second Edition. Chapman & Hall.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second Conference on Human Language Technology (HLT 2002)* San Diego, CA.

Doug Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. 2005. Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic. *International Conference on Intelligence Analysis*. McLean, VA.

Samuel Kotz, Norman Lloyd Johnson, and Campbell B. Read (Eds.). *Encyclopedia of Statistical Sciences*

Alon Lavie, Kenjie Sagae, and Shyamsundar Jayaraman. 2004. The Significance of Recall in Automated Metrics. In *Proceedings of the Association for Machine Translation in the Americas (AMTA 2004)* Washington, DC.

Chin-Yew Lin and Franz Josef 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 (ACL 2004)* Barcelona, Spain.

I. Dan Melamed, Ryan Green and Joseph P. Turian. 2003. Precision and Recall of Machine Translation Proteus technical report #03-004, a revised version of the paper presented at NAACL/HLT 2003, Edmonton, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)* Philadelphia, PA.

R Development Core Team 2004. R: A language and environment for statistical computing . *R Foundation for Statistical Computing*. http://www.R-project.org Vienna, Austria.

Kathryn Taylor and John White. 1998. Predicting What MT is Good for: User Judgements and Task Performance. . *Machine Translation and the Information Soup* Third Conference of the Association for Machine Translation in the Americas, AMTA Langhorne, PA, USA, pages 364-373. Springer, Berlin.

Clare R. Voss, Calandra R. Tate, and Eric Slud. 2004. Task-based Machine Translation Evaluation. *Presentation at MT Evaluation Panel, AMTA*. Georgetown University, Washington, DC.