Combining Evaluation Metrics Via Loss Functions

Calandra R. Tate[◇†]
[◇]Department of Mathematics
University of Maryland
College Park, MD 20742
ctate@math.umd.edu

Clare R. Voss[†]
[†]Army Research Lab
2800 Powder Mill Road
Adelphi, MD 20783
voss@arl.army.mil

Abstract

When response metrics for evaluating the utility of machine translation (MT) output on a given task do not yield a single ranking of MT engines, how are MT users to decide which engine best supports their task? When the cost of different types of response errors vary, how are MT users to factor that information into their rankings? What impact do different costs have on response-based rankings?

Starting with data from an extraction experiment detailed in Voss & Tate (2006), this paper describes three response-rate metrics developed to quantify different aspects of MT users' performance identifying who/when/where-items in MT output, and then presents a loss function analysis over these rates to derive a single customizable metric, applying a range of values to correct responses and costs to different error types.

For the given experimental dataset, loss function analyses provided a clearer characterization of the engines' relative strength than did comparing the response rates to each other. For one MT engine, varying the costs had *no* impact: the engine consistently ranked best. By contrast, cost variations *did* impact the ranking of the other two engines: a rank reversal occurred on who-item extractions when incorrect responses were penalized more than non-responses.

Future work with loss analysis, developing operational cost ratios of error rates to correct response rates, will require user studies and expert document-screening personnel to establish baseline values for effective MT engine support on wh-item extraction.

1 Introduction

Faced with foreign language (FL) texts that they cannot read, English-speakers want to know how effectively different machine translation (MT) engines will enable them to understand and make use of the information in those texts after translation. In 1998, Taylor & White proposed evaluating the utility of MT engines *extrinsically,* by measuring how accurately MT users could complete text-handling tasks on MT output. The tasks they proposed for the evaluation were presented in a hierarchy, to establish the ranking of MT engines that support those tasks. From their perspective, MT engines that support users performing linguistically more complex tasks should be ranked above (better than) MT engines that only support users on less complex tasks.[1] Their proposal set in motion task-based MT evaluation projects (Doyon et al., 1999; White et al., 2000; Laoudi et al., 2006; Voss & Tate, 2006) that have now raised new questions.

[1]Task-based evaluations measure discrete subject responses that can be scored against a ground truth set of answers, in contrast to subjective judgments of text translation quality that yield ordinal values. The relation between these two types of metrics has not been established. Whether there exists a statistically significant, predictive relation between (i) task-based metrics and subjective, text-based judgments of fluency and adequacy as Taylor & White suggested, or between (ii) task-based metrics and the automated text-based metrics proposed in the last few years—such as BLEU (Papineni et al. 2002), GTM (Melamed et al. 2003), METEOR (Lavie et al. 2004, Banerjee & Lavie 2005), and TER (Snover et al. 2005)—is an open research question. Tate (2005)'s early results modeling the relation in (ii) indicate that automated text-based metrics alone are not sufficient to predict task results.

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 242-250,
Cambridge, August 2006.

With the number and types of response metrics varying by task in the MT evaluation, the MT user may be faced with different MT engine rankings for each metric. How are MT users to decide which MT system best suits the given task in their work environment when there is no single, across-the-board ranking of MT engines? What method or methods of combining these metrics will be practical and interpretable by the MT user?

Furthermore, if the operational cost of different types of errors are not the same for the given task, how are the users to factor that information into their assessments? In different work environments, there may be also different values assigned to correct responses relative to the costs of missed or incorrect information. For example, in one environment with vast amounts of incoming, streaming data and substantial information redundancy across multiple channels, the cost of a particular missed item may be quite low relative to the value of correct detections. In another environment where data is limited, degraded, and slow to process, the costs of a missed or incorrectly categorized item may be substantially higher relative to the value of correct detections.

This paper proposes a compositional treatment of multiple cost-weighted, response-rates with loss functions to create a single metric for the purpose of ranking MT systems. Loss functions provide an established technique from statistical decision theory to fold in user-supplied costs that adjust the weight of distinct rates, to achieve an overall assessment of MT system quality, that can be treated as a figure of merit for ranking the systems in a particular environment. Our objective is ultimately to provide MT users with adequate methods for interpreting evaluation results for their tasks and work environment. Thus the costs applied to the response rates and the resulting analyses presented here are intended only as examples that the reader can work through in following the approach, rather than as actual operational values and outcomes.

The next section of the paper provides background for our starting analyses, with a brief description of the extraction task experiment and the raw response data collected in the experiment. In the sections that follow, we examine in more depth the definitional challenges inherent in developing response-rate metrics for evaluating all-occurrence extraction from MT output, and then we describe our approach applying loss functions to the derived response-rate metrics and present the results of loss function analysis on the provided dataset.

2 Background

The task data analyzed in this paper comes from a large-scale extraction experiment conducted at the University of Maryland[2] to assess three Arabic-English machine translation engines (Voss & Tate 2006). Before describing the experiment proper and the data collected as background, we clarify the relevant notions of extraction.

2.1 Extraction as a Task

If the users of a machine translation engine are analysts, for example, whose task objective is to find *all occurrences* of specific types of information, such as who/when/where-items, in the machine-translated text for later event interpretation and verification with drill-down requirements, then the "extraction task" most relevant to the Taylor & White hierarchy for evaluating MT engines is a text-markup or annotation process, identifying the locations of those items in the MT output. In evaluating task performance for a particular wh-item type, the annotations may be correct or incorrect responses, or the MT users may fail to detect some items (classed as a non-response)—yielding three basic performance events over which to define distinct extraction metrics.

If, on the other hand, the MT users are archivists who must find who/when/where-items in documents for later retrieval from a database, then the relevant "extraction task" for evaluating MT engines is more complex: it requires detecting items, determining their co-reference relations, and selecting or creating *one-best* items to go into slots in the database. The slot-responses may be filled correctly or incorrectly, the MT user may fail to detect some items, miss a slot, or create and fill spurious slots—yielding yet other basic performance events over which to define extraction metrics.

The distinction made here between *all-occurrence* and *one-best* extraction tasks is spelled out formally by De Sitter et al. (2004). For extensive

[2]The work was part of a research project for the Center for the Advanced Study of Languages (CASL), a university-affiliated research center at the University of Maryland.

details of one approach to evaluation that includes both these tasks and several others operating on entities, (numerical) values, temporal expressions, relations, and events, see the NIST guidelines and evaluation plans for automated content extraction (ACE 2006).[3]

2.2 Task-based Evaluation Experiment

The task-based approach to MT evaluation by Voss & Tate (2006) provides an experimental framework for evaluating subjects' performance on a *Who-*, *When-*, *Where-type* extraction task, given MT output. They conducted a large-scale experiment with fifty-nine subjects over two days, using a document collection and answer set that they constructed with ground-truthed Arabic texts and annotated English reference translations.

Via a web browser, each of the subjects in the experiment viewed 18 machine-translated documents, equally mixed across three Arabic-English MT systems and three Wh-types. For each machine-translated document that they viewed, subjects highlighted *all occurrences* or words or phrases that they identifed as being of the requested wh-type in the text displayed on the screen. Prior to the actual experiment, subjects were trained to identify who-, when-, and where-type elements in English-original and (Arabic-English) MT output texts. (Table 1 lists the categories included in each of the wh-types.) The task did not include mention linking for entity coreference or categorization of entity type as in ACE evaluations. During the training and evaluation phases of the experiment, their response selections were marked and stored directly in the text of copied files for later scoring.

2.3 Collected Dataset

The document collection for this experiment was drawn from Arabic news articles on websites of Arabic-speaking countries. The collection included six distinct articles for each of the three wh-types. Native Arabic speakers translated each text and identified the wh-items that served as ground truth for later determining the answer set in the MT output and establishing subject response types. All 18 Arabic source documents were then run through the three MT engines, yielding a full collection of 54 translated documents to be randomly distributed among subjects.[4]

Who-type: people, roles, organizations, companies, groups of people, and the government of a country
When-type: dates, times, duration or frequency in time, including proper names for days and common nouns referring to time periods
Where-type: geographic regions, facilities, buildings, landmarks, spatial relations, distances, and paths

Table 1: Wh-type Items in Extraction Task, from Voss & Tate (2006)

2.4 Basic Task Metrics

The basic level of aggregation of subject responses in the experiment was at the case-level, where a *case* was defined as one subject viewing one document as translated by one MT engine. First, three types of *event counts* for each of the experiment's 1060 cases were computed by comparing and identifying all of a subject's responses against all of the answer items in the translated document for that case as:

- *a correct response*, if a response fully matched an answer item, by covering all open class words, but possibly under- or over-extending with a determiner or other closed class word not crucial to the meaning of the wh-item
- *an incorrect response* if a response did not match any answer item in the translated document
- *a non-response* if no part of an answer item was marked by any of the subject's responses in the translated document.

For each translated document, the *total # answers* possible for the end-to-end evaluation was defined as the total # answer items in the reference translations, or RT total. For each case, the *total # subject*

[3]For further background on the origin of the U.S. government-sponsored, shared task evaluations that eventually led to the ACE program, see Sundheim (1989, 1996) and Palmer & Finin (1990).

[4]Details of the dataset construction steps and answer-set coding are described in Vanni, Voss, and Tate (2004).

responses was the count of subject-marked contiguous strings, and included fully correct, partially correct, and incorrect responses. Note that there was no given limit to the number of incorrect responses that a subject could produce in the process of extracting wh-items. So the subject-marked total (SubjMtot) varies by subject and document, and hence by case.

From these counts, three *event rate metrics* were computed over the cases for analyses at different levels of aggregation (such as by MT, by wh-type, and within wh-type by MT) as follows: *correct response rate* as #fully correct responses out of the RT total, *incorrect response rate* as #incorrect responses out of the subject-marked response total, and *non-response rate* as #non-responses out of the RT total.

3 Compositional Metrics

What are the desiderata in a compositional metric? For MT users, an MT evaluation methodology should provide an overall relative ranking of MT systems, where response rates can be combined somehow, with correct answers counting positively and errors counting negatively toward the overall performance of each MT system. In particular, a utility metric should penalize an MT system both when subjects fail to mark possible answer items and when they mark incorrect items.[5] Two types of compositional metrics already exist in the field of extraction research. However as briefly discussed below, neither suits the extraction task at hand.

3.1 F-Measure

De Sitter et al. (2004) review in formal detail the relation of extraction-type metrics—including the equivalents of correct responses (true positives), incorrect responses (false negatives), and non-responses (false positives)—to information retrieval metrics, such as precision, recall, and f-measure. Precision in the *all-occurrences* extraction corresponds to the proportion of fully correct responses out of all subject-marked responses (correct, partial, or incorrect), while recall corresponds to the proportion of correct responses out of all reference-translation (RT) answer items (correctly detected or missed as a non-response or lost in translation).[6] What their presentation makes clear however is that the f-measure, in composing precision and recall into one formula and thus bringing together all three basic metrics, does not allow for independently setting of the cost of different types of errors.

3.2 ACE Value Measures

By contrast, a quick look at the ACE (2006) guidelines and evaluation plan with its formulas for assigning value weights, value discounts, and other costs to entities for their type and attribute recognitions and for their mention detections, also makes clear that compositional metrics can become quite complex when numerous evaluation criteria are being judged.[7]

The extraction experiment that yielded the data at hand did not include the layers of analysis that the ACE tasks have. Subjects had no need to categorize the wh-items by type because for any one translated document that they saw, they only had to annotate one wh-type as specified on the screen. They had no need to track multiple mentions of the same entity because the task did not include coreference resolution. Thus, the simplicity of the experimental task pre-empted the need for as extensive a scoring manual as ACE has.[8]

[5] The latter plays the role of penalizing "guessing," as practiced in the scoring of educational and psychological tests.

[6] Since some RT answer items may be lost in translation by an MT engine and could not be detected by subjects, the total of the subjects' correct responses and non-responses must have that loss count added in to compute the full RT total for the denominator of the recall metric.

[7] Florian et al. (2004) found in analyzing their systems' performances in the ACE 2003 evaluation that the f-measure performance did not correlate well with improvements in their ACE value, because the weighting of entity types in the ACE formula had no corresponding adjustment in the f-measure. As a result, small improvements in the f-measure were paired with large relative improvements in the ACE value.

[8] Unlike the ACE conferences that evaluate only monolingual automated information extraction (IE), Babych and Hartley (2004) have reported evaluation results from automated named entity (NE) recognition *following* MT. They conclude, based on the DARPA-1994 MT data, that "for recognition of a large class of NEs, MT output is almost as useful as a human translation." Looking more broadly that NE recognition, Aone et al. (1997) concluded that information extraction (IE) preceeding MT yielded better results than the reverse sequencing of components with MT then IE. We note that it is beyond the scope of this paper to address these intriguing analyses further.

3.3 Loss Functions

The approach that we propose for combining the three response rates is to create a single loss function that incorporates all three rate types into one equation and estimates separate costs for each type. We follow (statistical) decision theory, as expounded in Raiffa & Shlaifer (1961) or Ferguson (1967), to estimate the expected recurring costs from different sources, on a per-instance basis, using a common monetary scale, as done in economics when forming a "utility function" or in game theory when defining a "payoff" function.

We imagine the future use of each MT system tested over a large series of N further document-screenings by human investigators without Arabic language training (as in this experiment). Each investigator tries to meet the information extraction goals like those in this experiment, with the goal of deciphering Arabic-language documents using machine-translated versions of the documents.

We assume that the individual (Document, Subject) combinations are sampled independently from the same universe of such combinations as was done in the experiment, for the fixed MT system. We also assume that costs are incurred, additively, (i) from the need for additional human screening of incorrectly marked documents and (ii) from the mitigation of wrong inferences drawn by human screeners from incorrect marks.[9]

We quantify

- the triple of average costs (-c1, c2, c3) where relative weights c1, c2, c3 > 0, in monetary units per occurrence, refer to correct response, non-response, and incorrect response items;
- the triple of rates (r1, r2, r3) = (correct response rate, non-response rate, incorrect response rate); and
- the average numbers respectively of RTtot, i.e., total count of RT answer items, and SubjMtot, i.e., total count of subject-marked items, per (Document, Subject) instance.

[9] Automatic MT evaluation metrics could also, but currently do not, penalize errors differentially: why not have higher costs, for example, on open-class word translation errors than closed class translation errors (Calison-Burch et al. 2006)? This would be consistent with scoring schemes applied to human translations (Kovarik 2005).

	MT 1	MT 2	MT 3
Correct Responses (CR)	1181	1506	1370
Non-Responses (NR)	558	573	585
Incorrect Responses (IR)	438	311	513
Total RT Answers (RTtot)	3091	3066	3086
Total Subject Marks (SubjMtot)	2759	2636	2842
Correct Response Rate (CRR)	.382	.491	.444
Non-response Rate (NRR)	.181	.187	.190
Incorrect Response Rate (IRR)	.159	.118	.181

Table 2: Response counts and rates by MT

Since each MT system would be used repeatedly on independently generated instances, the costs or "loss" resulting from a specific number of correct response items, incorrect response items, and non-response items in each (Document, Subject) would be summed over N distinct instances and would (because of the Law of Large Numbers) yield an expression roughly equal to N multiplied by the per-instance average-loss expression

Average Loss = (-c1 * r1 + c2 * r2)* E(RTtot)
+ (c3 * r3) * E(SubjMtot)

This expected-loss function, while simple and reasonable, is by no means the only possible one, but is particularly tractable. Its linear form arises from the assumption that gains or losses of different categories of mark are additive; the way in which rates and expectations enter the formula are otherwise a consequence of the Law of Large Numbers and linearity properties of Expectation.

4 Results and Analyses

4.1 Basic Metrics: Response Counts and Rates

The response counts and response rates are presented by MT and by Wh-type x MT category in Tables 2 and 3 respectively. From the response rates in Table 2, it is clear the MT2 yields the subject performance that is best (lowest) for incorrect response rate and best (highest) for correct response rate on this task. Statistically, while there is a significant difference between MT2 and the other MT engines on the IRR, there is however no significant difference on the CRR for MT2 and MT3, and there is no significant difference among the engines on their non-response rates.[10] While the data in this table

[10] Voss & Tate (2006) report details of statistically significant effects in chi-square tests of equality for CRR and IRR over MT, and for NRR and IRR over Wh-x-MT.

Wh	MT	CRR	NRR	IRR
When	1	.333	.218	.148
Where	1	.387	.211	.173
Who	1	.417	.120	.154
When	2	.474	.178	.127
Where	2	.515	.214	.145
Who	2	.481	.167	.087
When	3	.410	.216	.173
Where	3	.443	.207	.173
Who	3	.472	.151	.193

Table 3: Response rates by Wh-x-MT category

	MT1	MT2	MT3
E(RTtot)	8.73	8.69	8.74
E(SubjMtot)	7.79	7.47	8.05

Table 4: Expected totals by MT system

clearly indicates that MT2 is the leading contender for top rank, the developers of the MT3 engine can point out there is no statistically significant advantage for MT2 on CRR and NRR over their engine. If incorrect responses were not penalized, they could argue that their engine MT3 should be ranked alongside MT2.

Table 3 shows the response rates cross-classified by Wh x MT, providing for finer-detailed data analysis. Could it be that one MT yields better across-the-board response rates for one Wh-type extraction, while another MT is better for another Wh-type? For IRR and NRR, however, this is not the case. On who-items, for example, MT1 yields significantly better (lower) NRR than MT2 and MT3, while MT2 is better (lower) on IRR than the other two systems. With the detailed breakout of the data, the factors to consider in ranking of the engines become more complex.

4.2 Compositional Metrics: Loss Function Analyses

To work with the Average Loss expression, we estimate the rates and expectations from corresponding quantities in the experiment data. For costs, however, we can at this point only "guess" and select hypothetical values as placeholders, for the purpose of showing how the method of applying a loss function works. Note that the rates and expectations are specific to the experiment's MT systems under consideration, but the weights (c1, c2, c3) are not.

	MT1	MT2	MT3
Average Loss	-12.30	-17.12	-14.60

Table 5: Average Loss within MT System with user-specified costs (-c1,c2,c3)=(-5,2,1)

	MT1	MT2	MT3
Average Loss	2.30	0.75	2.35

Table 6: Average Loss within MT System with user-specified costs (-c1,c2,c3)=(-1,2,2)

Case 1

The values of the rates and expectations estimated from the current data, as a function of MT, are presented in tables 2 and 4 respectively. The values of the cost weight-parameters, only the ratios of which really matter to the ranking by Average Loss of the MT systems, we take to be: (-c1, c2, c3)= (-5, 2,1).

The rationale for the selected values is that the importance of *correct responses* clearly outranks the importance of errors, in the form of *non-responses*, which in turn may be more costly on a per-item basis than *incorrect responses*. The Average Loss numbers calculated by the formula above are shown in Table 5 with an overall preference ranking of MT systems, where a lower cost value is better "$>$" than a higher cost value, as: MT2 $>$ MT3 $>$ MT1.

We tried numerous combinations of c1, c2, and c3 values, but none dislodged MT2 from its rank as best. If however we were to penalize incorrect responses and non-responses equally and (perhaps unreasonably) at twice the weight of correct responses, with (-c1, c2, c3)= (-1, 2, 2), then as shown in Table 6, the ranking of MT1 and MT3 can be tied.

To cause MT1 to outrank MT3, significantly unbalanced costs are needed. As Table 7 indicates, with the severity of the penalty on non-responses increased even further, with (-c1, c2, c3)= (-1, 5, 2), MT1 can outrank MT3 a bit more, while MT2 retains its rank as first.

Case 2

We recognize that, in some work environment, a particular number of non-responses might deserve as much as or even more weight than a single correct response. While we cannot hope to be precise in

	MT1	MT2	MT3
Average Loss	7.04	5.62	7.34

Table 7: Average Loss within MT System with user-specified costs (-c1,c2,c3)=(-1,5,2)

		MT1	MT2	MT3
E(RTtot)	When	7.47	7.48	7.51
	Where	9.38	9.27	9.35
	Who	9.35	9.30	9.36
E(SubjMtot)	When	5.90	6.11	6.90
	Where	7.66	7.80	8.19
	Who	9.82	8.48	9.06

Table 8: Expected totals by MT system and WH type

Average Loss	MT1	MT2	MT3
When	-8.31	-14.29	-10.96
Where	-12.87	-18.77	-15.42
Who	-15.74	-18.52	-17.51

Table 9: Average Loss within Wh-type by MT System with user-specified costs (-c1,c2,c3)=(-5,2,1)

Average Loss	MT1	MT2	MT3
When	.51	-6.80	-0.72
Where	-0.97	-7.73	-3.24
Who	-5.57	-10.54	-5.41

Table 10: Average Loss within Wh-type by MT System with user-specified costs (-c1,c2,c3)=(-5,5,5.5)

specifying these costs, we can provide some alternative MT-system rankings based on reasonable hypothetical values.[11] For instance, an alternative calculation of losses could be made within specific WH contexts for each MT system.

We illustrate here that the ranking of MT systems is uniformly MT2 $>$ MT3 $>$ MT1 within specific WH-categories.[12] The average loss results for the MT system by Wh-type combinations are based on calculations with rate values from Table 3 above and expected values for SubjMtot and RTtot in Table 8.

Table 9 shows the Average Loss within Wh-type by MT System. As can be seen by comparing the rankings within rows of this table, the ranking MT2 $>$ MT3 $>$ MT1 persists even within each of the three Wh-types using the same cost-weights (-c1, c2, c3)= (-5,2,1).

Case 3

The same analysis but where costs per correct response are adjusted to -c1=-5, c2=5, and c3=5.5, as shown in Table 10, for the corresponding estimated rates and expected totals from Tables 3 and 8 respectively. When shifting the weights to emphasize more heavily the cost of erroneous marks, setting c1=c2$<$c3, we see that the overall ranking

$$\text{MT2} > \text{MT3} > \text{MT1}$$

persists, except for a reverse of the ranking between MT1 and MT3 within the Wh=Who category.

Other Cases

Dozens of other combinations of weights with the cross-classified WhxMT data were also tried (including c1$<$c2$<$c3 and c1$<$c3$<$c2) but we found that, for reverse ranking of MT1 and MT3 to be substantial, correct responses would have to be considerably discounted relative to the other response types, which does not seem to be a desirable objective. Furthermore, even with the modified weights and various combinations tested, the previous MT rankings persist in all other categories, and MT2 still retains its superiority within Wh-type.

5 Summary and Future Work

We applied an approach from statistical decision theory, combining three performance metrics—correct response, non-response, and incorrect response rates—with weighted cost estimates into a single *loss function*. Based on several tests with hypothetical costs, the overall preference ranking of MT systems placed MT2 as front runner, while either other MT engine could be ranked next, based on assigned response costs.

Longer term, we seek to support our users in determining which MT system (or possibly which combination of MT systems) best meets the cost and accuracy requirements of their tasks. This will entail working with expert document-screening personnel to develop cost ratios of incorrect-response and

[11] This approach will effectively yield clearer and more meaningful MT-system rankings when the relative-importance ratios of c1/c2 and c1/c3 are elicited from experts.

[12] Voss & Tate (2006) report that chi-squared tests for "interaction" of response between the MT and Wh classifications, with respect to response-rates, show no significant interaction for differences in correct response rates (chi-square 8.96, p-value .061 on 4 degrees of freedom) and highly significant difference in both incorrect response rates and non-response rates (chi-square 16.45 and 15.17 respectively, with p-values .002 and .004 on 4 degrees of freedom).

non-response error rates to correct-response rates, based on their cumulative experience with foreign-language documents, and then comparing costs as calculated with loss functions against empirical results from user studies.

Acknowledgements

Several individuals contributed to the task-based evaluation research project, including Eric Slud (Dept. of Mathematics, U. of Maryland, College Park), Matthew Aguirre, John Hancock (ArtisTech, Inc.), Jamal Laoudi, Sooyon Lee (ARTI), and Somiya Shukla, Joi Turner, and Michelle Vanni (ARL). This project was funded in part by the Center for Advanced Study of Language (CASL) at the University of Maryland.

References

ACE 2006. The ACE 2006 Evaluation Plan:. Evaluation of the Detectin and Recognition of ACE. Entities, Values, Temporal Expressions, Relations, & Events. *www.nist.gov/speech/tests/ace/doc/ace06/index.htm*.

C. Aone, H. Blejer, M.E. Okurowski, C. Van Ess-Dykema. 1994. A Hybrid Approach to Multilingual Text Processing: Information Extraction and Machine Translation. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas* (AMTA-94). Columbia, Maryland.

B. Babych & A. Hartley. 2004. Comparative Evaluation of Automatic Named Entity Recognitioin from Machine Translation Output. In *Proceedings of Workshop on Named Entity Recognition for Natural Language Processing Applications*. Held in conjunction with with the First International Joint Conference on Natural Language Processing (IJCNLP-04). Sanya, Hainan.

S. Banerjee & A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlations with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005). Ann Arbor, Michigan.

C. Calison-Burch et al. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. *Proceedings of EACL*. Genoa, Italy.

A. De Sitter, T. Calders, W. Daelemans. 2004. A Formal Framework for Evaluation of Information Extraction. Technical Report TR 2004-0, Dept. of Mathematics and Computer Science, University of Antwerp, Belgium.

J. Doyon, K. Taylor, & J. White. 1999. Task-based Evaluation for Machine Translation. In *Proceedings of Machine Translation Summit VII*. Singapore.

T. S. Ferguson. 1967. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press. NY.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, J. Kambhatla, X. Luo, N. Nicolov, & S. Roukos. 2004. A Statistical Model for Multilingual Entity Detection and Tracking. In *Proceedings of HLT/NAACL*. Boston, MA.

V. Goel & W. Byrne. 1999. Task Dependent Loss Functions in Speech Recognition: Application to Named Entity Extraction. In *Proceedings of the ESCA ETRW Workshop, Accessing Information in Spoken Audio*. Cambridge, UK.

J. Kovarik. 2005. Presentation, *Guidelines for Scoring Human Translations*. DARPA GALE PI meeting. San Jose, CA.

A. Lavie, K. Sagae, & J. Jayaraman. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA-2004). Washington, DC.

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

M. Palmer & T. Finin. 1990. Workshop on the Evaluation of Natural Language Processing Systems. *Computational Linguistics*. 16(3).

K. Papineni, S. Roukos, T. Ward, & W. Zhu. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the ACL*. Philadelphia, PA.

H. Raiffa and R. Schlaifer. 1961. *Applied Statistical Decision Theory*. Harvard Business School. Boston, MA.

M. Snover, B.J. Dorr, R. Schwartz, J. Makhoul, L. Micciulla, & R. Weischedel. (2005). A Study of Translation Error Rate with Targeted Human Annotation. LAMP-TR-126. U. of Maryland, College Park.

B. Sundheim. 1989. Plans for a Task-Oriented Evaluation of Natural Language Understanding Systems. In *Proceedings of DARPA Speech and Natural Language Systems Workshop*. Philadelphia, PA.

B. Sundheim. 1996. Ch 13.2 Task-Oriented Text Analysis Evaluation. In *Survey of the State of the Art in Human Language Technology*. Ronald A. Cole, Editor in Chief, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, Victor Zue. Sponsored by the National Science Foundation and the European Commission.

K. Taylor & J. White. (1998). Predicting What MT is Good for: User Judgements and Task Performance. *Proceedings of AMTA*, pages 364–373. Springer, Berlin.

C. Tate. (2005). Evaluating Machine Translation Output & Predicting Its Utility. *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* . Ann Arbor, MI.

M. Vanni, C. Voss, and C. Tate. 2004. Ground Truth, Reference Truth & Omniscient Truth: Parallel Phrases in Parallel Texts for MT Evaluation. In *Proceedings of LREC*. Lisbon, Portugal.

C. Voss and C. Tate. 2006. Task-based Evaluation of Machine Translation (MT) Engines. Measuring How Well People Extract Who, When, Where-Type Elements in MT Output. In *Proceedings of the EAMT*. Oslo, Norway.

J. White, Doyon, J., & Talbott,S. 2000. Task Tolerance of MT Output in Integrated Text Processes. In *Proceedings of Workshop: Embedded Machine Translation Systems*. ANLP/NAACL. Seattle, WA.