Task-based Evaluation of Machine Translation (MT) Engines: Measuring How Well People Extract Who, When, Where-Type Elements in MT Output

Clare R. Voss[◇], **Calandra R. Tate**[♠◇]

[◇]Multilingual Computing Group, ARL, Adelphi, MD 20783 (USA)
[♠]Dept. of Mathematics, University of Maryland, College Park, MD 20740 (USA)
voss@arl.army.mil, ctate@math.umd.edu

Abstract

How effectively can people perform the text-handling task of extracting information from the output of MT engines? When is the output of one MT engine more likely than the output of another engine to support people performing an extraction task? This paper reports on the results of a one-of-a-kind, large-scale, MT evaluation experiment where nearly sixty subjects extracted who, when, and where-type elements of information (EIs) from output generated by three types of Arabic-English MT engines. Our hypothesis was that, in an end-to-end (MT engine and user) evaluation, the best extraction results would come from subjects working with output from MT engines that reordered Arabic input to generate English word order, rather than from an engine that did not, i.e., from a statistical or a rule-based, rather than from a substitution-based MT engine.

The results of the experiment were not as straight-forward as expected: (1) *non-response rates* were statistically comparable across all three evaluated MT engines, while (2) *correct response rates* were statistically comparable on two engines, the statistical and substitution-based engines that yielded better (higher) rates than the rule-based engine did, and (3) *incorrect response rates* were statistically comparable on a different pair of two engines, the rule- and substitution-based engines that yielded significantly worse (higher) rates than the statistical engine. While these results do indicate that the statistical engine yielded significantly better rates than at least one of the other two engines on two of the three metrics, the *lack of uniform results* pre-empts an across-the-board ranking of the engines.

Our next step is to incorporate the collected data in statistical models and test for their adequacy in predicting these task results from faster and less expensive, automatic metrics. The long-term goal is to understand which metrics accurately predict MT users' task effectiveness with different MT engines on text-handling tasks of varying levels of difficulty.

1 Introduction

Among machine translation (MT) developers for over a decade, there has been the assumption that MT engines are "good enough" to support people performing certain applications in the real world (Church & Hovy 1993). More recently, informal reports from operational and field settings have described successful, but carefully limited use of MT output in real-world tasks. (Fisher, et al. 1999; Holland, 2005). The research reported here was undertaken to assess how effectively people can perform one specific real-world task on the outputs of different MT engines. The paper describes the results of *a one-of-a-kind, large-scale, MT evaluation experiment* where nearly sixty subjects extracted who, when, and where-type essential elements of information (EIs) from output generated by three types of Arabic-English MT engines.

Our hypothesis was that, in an end-to-end (MT engine and user) evaluation, the best extraction results would come from subjects working with output from MT engines where Arabic phrases are explicitly re-ordered for English translation (rather than from an engine that does not reorder), i.e., from a statistical or a rule-based MT engine rather than from a substitution-based MT engine.

The statistically significant results of the experiment were not as uniform as expected. While the statistical engine yielded the best (lowest) *incorrect response rates,* it yielded results comparable to those of the substitution-based engine on *correct response rates,* and comparable to both the substitution-based and rule-based engines on *non-response rates.* This lack of uniform results pre-empts an across-the-board ranking of the MT engines. To assist users in interpreting these results, we are now testing the use of loss functions that will yield a single customizable metric from the three rates, based on acceptable costs for errors and failed detections, as provided by MT users for their work environment.

Our next step is to incorporate the collected data in statistical models and test for their adequacy in predicting these task results from faster and less expensive, automatic metrics. Our long-term goal is to understand which metrics accurately predict our MT users' task effectiveness with different MT engines on text-handling tasks of varying levels of difficulty.

This paper describes the background and motivation for our selection of an extraction task and a particular set of three MT engines. An overview of the experiment follows with a description of the task and elements to be extracted, the document collection, the subjects, the experimental design and procedures, and the data collected. The metrics applied to the data are detailed and the results and analysis of the experimental data are presented. The paper concludes with a recap of our findings and a few words about future directions for our work.

2 Background

Extrinsic, task-based evaluation of MT engines has long been of interest to MT users who seek automated support tools to expedite their decision-making tasks (Spaerck-Jones & Gallier, 1996). In the late 1990's two new MT research trends emerged, furthering interest in extrinsic metrics: task-based experiments were being conducted by MT developers on their own engines (Resnik, 1997; Levin et al., 1999), and task-based experiments assuming an ordering of task difficulty were being proposed by users for text-handling tasks (Taylor & White, 1998; White et al., 2000).

Then, with the introduction of several automatic MT metrics[1] demonstrating both the vitality of MT evaluation as a research area of its own and the impact of metrics on the MT development cycle, MT stakeholders began funding research experiments in task-based assessment of MT engines, to address users' needs. See, for example, the request for proposals that include methods for *utility evaluations,* in the 2005 broad agency announcement for GALE (Global Autonomous Language Exploitation), a large research program, directed by DARPA, a U.S. government funding agency.[2]

Publishing	Produce technically correct document in fluent English
Gisting	Produce a summary of the document
Extraction	For documents of interest, capture specified key information
Deep Extraction	Event identification (scenarios): determine an incident type and report all pertinent information
Intermediate Extraction	Relationship identification: member-of, friend-of, boss-of
Wh-Item Extraction	Identification of: who-, where-, when-type information elements
Shallow Extraction	Named entity recognition: isolate names of people, places, organizations, dates, locations
Triage	For documents of interest, rank by importance
Detection	Find documents of interest
Filtering	Discard irrelevant documents

Table 1: Task Hierarchy by Taylor & White (1998) with extra row inserted for Wh-item Extraction Task

Tasks for MT Evaluation

After reviewing Taylor & White's hierarchy of tasks and examining the MT output of several engines, we designed three experiments to test for:

(i) one task as a lower-bound for a shared capability, that all the selected different types of Arabic-English MT engines could support,

(ii) one task as an intermediate challenge, that one or two engines would support but another one would not, and

(iii) one task as an upper-bound for a shared limitation, that none of the selected engines could yet support.

This report focuses exclusively[3] on the

[1] Such as BLEU (Papineni et al. 2002), GTM (Melamed et al. 2003), METEOR (Lavie et al. 2004), and TER (Snover et al. 2005).

[2] www.darpa.mil/ipto/solicitations/closed/05-28_PIP.htm

[3] The choice for task (i) as topic categorization, a form of Taylor & White's "detection," and for task (iii) as template completion, a form of Taylor & White's "deep extraction" were based on our previous experiments (Tate, Lee, & Voss, 2003; and Voss, 2002). The results of the pilot conducted for task (iii) indicated that the MT engines were not adequate to support users performing event-template completion (Laoudi, Tate, & Voss, 2006).

experiment conducted for task (ii), that we gauged to be at an intermediate level of extraction difficulty, shown as an extra row inserted in Table 1. A small, prior pilot experiment to evaluate Arabic-English MT engines for document exploitation tasks indicated that subjects could extract some named entities and event participants from noisy MT output, but they could not readily identify relations within events (Voss, 2002). This led us to select wh-item extraction, a task between event-level analysis and named-entity recognition.

Selection of MT Engines

In conjunction with the project sponsor, three distinct types of MT engines were selected[4] as representative of three development models, varying in required funding, time, and linguistic resources:

- MT-1, a rule-based engine with hand-crafted lexicons and symbolic linguistic processing components (morphological analyzer, parser)
- MT-2, a statistical engine trained on large quantities of monolingual and parallel Arabic-English texts, but no traditional symbolic linguistic processing components
- MT-3, a substitution-based engine that relies entirely on a pattern-matching algorithm with a lexicon and morphological analyzer to translate matched strings into English phrases, replacing the former with the latter, leaving the original Arabic word order unchanged except as occurs locally within the substituted phrases.

3 Experiment

Here we provide an overview of the experiment with a brief description of the document collection, the experimental design, the task and wh-type elements to be extracted, and the data collected.

Document Collection

A collection of Arabic news documents taken from ten websites was created in December 2003. Full articles were trimmed from the bottom up to be roughly comparable in size and fit fully within the software display window after translation, so that subjects would not need to use a scrollbar to see any portion of the text.

X	Y	Z
MT-1 Who 1	MT-3 Who 1	MT-2 Who 1
MT-3 When 1	MT-1 When 1	MT-2 When 1
MT-1 Where 1	MT-2 Where 1	MT-3 Where 1
MT-3 Who 2	MT-1 Who 2	MT-2 Who 2
MT-2 When 2	MT-3 When 2	MT-1 When 2
MT-2 Where 2	MT-3 Where 2	MT-1 Where 2
MT-1 When 3	MT-2 When 3	MT-3 When 3
MT-3 Where 3	MT-1 Where 3	MT-2 Where 3
MT-2 Who 3	MT-3 Who 3	MT-1 Who 3
MT-3 When 4	MT-1 When 4	MT-2 When 4
MT-3 Who 4	MT-2 Who 4	MT-1 Who 4
MT-3 Where 4	MT-1 Where 4	MT-2 Where 4
MT-1 When 5	MT-2 When 5	MT-3 When 5
MT-2 Who 5	MT-1 Who 5	MT-3 Who 5
MT-1 Where 5	MT-2 Where 5	MT-3 Where 5
MT-2 When 6	MT-3 When 6	MT-1 When 6
MT-1 Who 6	MT-2 Who 6	MT-3 Who 6
MT-2 Where 6	MT-3 Where 6	MT-1 Where 6

Table 2: Super-Block of Three Viewing Sequences of 18 Translated Documents Each, Before Wh-type Grouping and Randomizations. Each translated document in the full pool of 54 is uniquely specified by an MT identifier (MT-1, MT-2, or MT-3), a WH identifier (WHEN, WHERE, or WHO), and a DocID identifier (an integer from 1 to 6).

For each of the three wh-item types, native Arabic speakers identified six different trimmed documents with between six and ten wh-items of that type in the text. The documentation of these wh-items in the resulting 18 Arabic document collection, established the ”ground truth” (GT) items for later determining the experiment’s answer set. All 18 Arabic source documents were then run through the three MT engines, yielding a full experiment collection of 54 translated documents.

Experimental Design

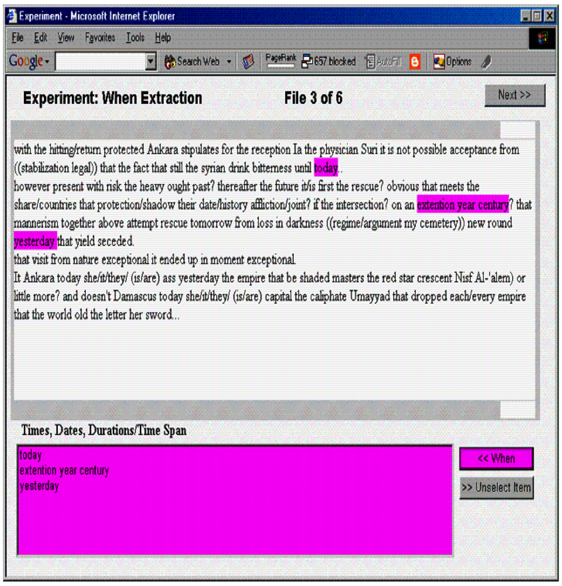
The design for the experiment assumed 60 subjects total. Each subject was assigned a pre-arranged, randomized sequence of 18 documents out of the full pool of 54 machine-translated documents. The sequences were constructed as follows. First one super-block was filled with three viewing sequences that included the entire 54-document set, as shown in Table 2. As a result, with three subjects, all the translated documents were viewed exactly once.

[4] The developers of the selected engines provided us with their most recently released version in November 2003, so that we could prepare for the experiments to be conducted in January 2004.

The following set of four constraints on the experimental design ensured a balance of the number of MT and Wh-type viewings per subject across the document collection.

1. No subject saw the same document more than once, i.e., as translated by more than one MT system.
2. Each subject viewed translated documents from each system an equal number of times.
3. Each subject viewed translated documents from each of the three Wh-type an equal number of times.
4. Each subject viewed the documents of each Wh-type one after another, without shifting back and forth among the three Wh-type groups.

After filling the super-block to meet these constraints, the next steps were randomizing the elements within each block while preserving the wh-type grouping of constraint 4. within each viewing sequence of the super-block, and then including enough replications for all subjects. Randomizing was done to prevent bias. Two super-blocks were randomized individually. Then the resulting 6-block randomization was replicated 10 times for the proposed 60-person full experiment.



Experiment: When Extraction

File 3 of 6

with the hitting/return protected Ankara stipulates for the reception Ia the physician Suri it is not possible acceptance from ((stabilization legal)) that the fact that still the syrian drink bitterness until today..
however present with risk the heavy ought past? thereafter the future it/is first the rescue? obvious that meets the share/countries that protection/shadow their date/history affliction/joint? if the intersection? on an extention year century? that mannerism together above attempt rescue tomorrow from loss in darkness ((regime/argument my cemetery)) new round yesterday that yield seceded.
that visit from nature exceptional it ended up in moment exceptional.
It Ankara today she/it/they/ (is/are) ass yesterday the empire that be shaded masters the red star crescent Nisf Al-'alem) or little more? and doesn't Damascus today she/it/they/ (is/are) capital the caliphate Umayyad that dropped each/every empire that the world old the letter her sword...

Times, Dates, Durations/Time Span

today
extention year century
yesterday

Figure 1: Screen shot of When-extraction (sample subject responses highlighted in upper text box & listed in lower answer box

Wh-item Extraction Task

From the outset, we knew that subjects would be available for most of one working day to participate in the three project experiments. This limited the time to conduct the training, practice, and evaluation phases of task (ii) to roughly two hours. We designed software to enable subjects to view all documents for all these phases via a browser and simply to click or click-and-drag over text they selected as wh-items that would then appear in an answer box at the bottom of the screen (see Figure 1).

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 3: Wh-type Items in Extraction Task

To determine how best to train people quickly and accurately on identifying who-, when-, and where-type elements in MT output (see Table 3), we first conducted two rounds of pre-piloting with English documents (no MT) where English speakers were given different definitions of the elements of information to highlight with colored markers on hard-copy English documents.

We conducted one pre-pilot with hard-copy MT output and then one 12-person pilot over the internet via browsers, where the instructions were augmented and refined with examples to clarify what would count as an element in non-fluent, i.e., MT output, text. We found that people were most thorough and consistent when instructed: (a) to scan for all cases of *only one wh-type* in a document and (b) to spot for clear case words of the wh-type they sought at first, and then to include other words adjacent to the clear case, as long as the resulting longer string remained of the same wh-type.[5]

As a result, in the final experiment, subjects viewed documents of each wh-type one after another, without shifting back and forth among the three wh-types, and the screen display in-

[5] For example, after first spotting *airport* as a where-type, the subject could scan and extend the item to its full PP *outside the airport,* because that form would provide more information for an accurate identification of the cited location.

dicated which wh-type to extract for a given document. For example, in Figure 1 the subjects were only to extract When-type items from the translated document on the screen. Thus, unlike the automated extraction tasks of evaluation programs such as ACE,[6] our extraction task only requires selecting text: there are no additional steps for categorizing the text (since the wh-type is given) or detecting co-reference among the the extracted elements.

The full evaluation experiment was conducted on two days, with 30 subjects participating on day 1 and another 29 subjects on day. The experiment was monitored by several observers and the software, run from an off-site server, was controlled by an administrator who monitored the subjects' progress online in real-time during the experiment, as they sat in the same large teaching classroom at individual computer workstations.

Subjects were instructed together on the task during the training phase by one project investigator. They received hard-copy pages with definitions and example wh-type items that they were told they could refer to at any phase in the experiment. During the practice phase before the actual experiment tests subjects worked through the extraction process on 9 documents (1 original English and 2 translated English for each of the 3 wh-types), with no spoken instruction. All subjects practiced the task on the same sequence of documents and received the same feedback with correct responses and brief descriptions via their browsers, following their responses.

Data Collected

In the full experiment, with 59 subjects each viewing 18 translated documents (arranged to include 6 documents from each of the 3 MT engines and 6 from each of the 3 wh-types), we collected 1062 *cases* (where one subject responded to one document translated by one MT engine).[7] However, with one server-client connection crash on day 2, two instances of translated documents viewed and marked by subjects could not be processed. The final data analyses were run on 1060 cases. With more than 100 wh-items total in the 54 translated collection, we collected well over 10,000 subject-extracted items to be evaluated.

4 Metrics

Answer Set for Full Document Collection

The creation and validation of the wh-item answer set against which the subjects' responses were scored, was constructed in three stages. Here we only summarize briefly the process involved. For a full explanation of the details, see Vanni et al. (2004).

First, *ground-truth wh-items* (GT) were identified in the original Arabic documents by a native Arabic speaker who placed them in an inventory spreadsheet, with one item per row. The original documents were also fully translated by each of four human translators working on the project, providing reference translations for later research comparing task-based results with automated metric scores.

In the second stage, the lead translator and one professionally trained native-English linguist worked together with the resulting reference translations and the GT-annotated original documents. They identified the *reference-truth wh-items* (RT) in the reference translations and placed each of these alongside the corresponding GT item in the inventory spreadsheet.

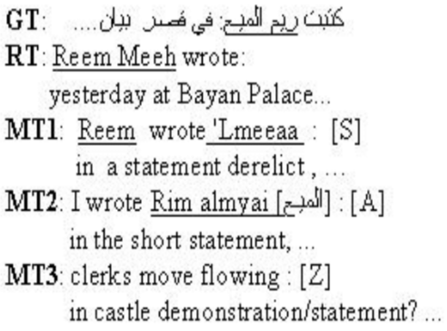
In the final stage, six individuals independently examined pairs of the reference translations and MT outputs side-by-side, and recorded into their separate inventory spreadsheets, the *omniscient-truth wh-items* (OT) in the MT outputs, i.e., those strings of words found to correspond to the RT items. Figure 2 shows one example of underlined phrases placed in correspondence to each other during the coding process: the GT item in the source text, the RT item in the reference translations, and the OT items in the MT outputs.

After several rounds of analysis, we developed a stable set of annotation instructions for defining the OT identification process in MT output text. Annotators both identified OT items and categorized them, as A if accurate semantically and grammatically, as B if either semantically or grammatically incorrect or incomplete, as S if the phrase was split across other content words that were not of that wh-type, and as Z if the item could not be identified in the MT output or was lost in the translation. When the inter-

[6] Evaluations under the Automatic Content Extraction program are described at www.nist.gov/speech/tests/ace/.

[7] The study had 59 subjects because one subject of the originally expected 60 did not show.

annotator agreement rates were acceptably high on the annotation process, we finalized the selection of the OT items for each MT engine's output.



GT: ....كتبت ريم المبح: 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 [المبح] : [A]
in the short statement, ...

MT3: clerks move flowing : [Z]
in castle demonstration/statement? ...

Figure 2: Sample Parallel Phrases in Parallel Texts: Who ground truth (GT) phrase in Arabic source text, Who reference truth (RT) phrase in reference translation, and Who omnisicient truth (OT) phrases in MT1, MT2, and MT3 output. [A],[S],and [Z] are OT-item annotation codes.

The annotation of the GT, RT, and OT items across the Arabic texts, the reference translations, and the machine-translated texts respectively, produced two types of extraction-task answers used in subsequent computations. Note: we distinguish these "answers" described here as established in documents before the experiment from the "responses" that subjects provide in the experiment proper.

1. the set of RT items
   - they are independent of the MT engines
   - they were defined in one-to-one correspondence with the GT items, and so are the English equivalents of the essential elements of information in the Arabic texts
   - they are used as denominators to calculate recall metrics (see next section)

2. the set of OT items
   - they are MT engine-specific
   - they were defined and annotated in correspondence to the RT items, and so are the machine-translated equivalents of the essential elements of information in the Arabic texts
   - they are compared in automatic text-matching algorithms against the subjects' responses because they are the content of the essential elements as made available to the subjects by each MT engine

For each Arabic document in document collection and for each machine-translated version of that document, the set of RT items was fixed. By contrast, for each machine-translated version of each Arabic document, the set of OT items could and often did vary both in content of the translated item and in number of translated items (because some were lost in translation).

Task Metrics

Three task metrics for evaluation were computed as follows. First we tallied three types of *event counts* for each of the 1060 cases, by comparing and identifying all of a subject's responses against all of the (OT) 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 # RT items in the reference translations, or RT total. For each case, the *total # subject 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 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.

	MT 1	MT 2	MT 3
# correct-responses	1181	1506	1370
# non-responses	558	573	585
# incorrect-responses	438	311	513
total # RT items	3091	3066	3086
total # Subj resp.	2759	2636	2842
Correct-response rate	.382	.491	.444
Non-response rate	.181	.187	.190
Incorrect-response rate	.159	.118	.181

Table 4: Event counts and rates tallied over translated-document by subject cases within each of three MT systems

	When	Where	Who
# correct-responses	1068	1480	1509
# non-responses	538	696	482
# incorrect-responses	334	456	472
total # RT items	2635	3304	3304
total # Subj resp.	2218	2790	3229
Correct-response rate	.405	.448	.457
Non-response rate	.204	.211	.146
Incorrect-response rate	.151	.163	.146

Table 5: Event counts and rates tallied over translated-document by subject cases within each of three Wh-types

5 Results and Analyses

We tallied the event counts and computed their corresponding rates by MT in Table 4, by Wh-type in Table 5 and by Wh-by-MT in Table 6.

Event Results by MT Engine

In Table 4 we tally the correct responses, non-responses, incorrect responses as well as the RT totals[8] and subject-marked totals by MT system. In addition, we present the overall fully correct response rates, the non-response rates, and the incorrect response rates for each MT system.

While the subjects' *non-response rates* in extracting wh-items from MT output were comparable across three MT engines (statistically indistinguishable), one engine yielded significantly weaker (lower) *correct-response rates,* and another engine yielded significantly strong (lower) *incorrect-response rates.* In particular, a chi-square test for equality of correct response rate and non-response rate over MT yield wildly significant statistics (each on 2 degrees of freedom) respectively equal to 74.89 and 42.19.

Given MT-1's syntactic analyses and MT-2's phrase-based modeling, we had expected stronger subject performances on their MT output (where phrases are re-ordered for English), than on the output of MT-3's substitution-based translation (where Arabic syntax is *not* changed to generate English word order). That is, we hypothesized that subjects would show better task performance on MT-1 and MT-2 than MT-3. However, the statistical analyses of subjects' extractions on this task however did *not* match these predictions.

Event Rates by Wh-type

Table 5 displays tallies and rates of events (the same ones as in Table 4) within all translated documents of each Wh-type, without regard to MT engines. In this table, there are highly significant differences among the correct response rates and the non-response rates for different Wh-types of documents (with chi-square statistics respectively 17.43 and 54.20 on 2 degrees of freedom): clearly the correct response rates are

[8] We note here for clarification that the RT totals do not match across MT engines because we experienced two types of document losses. Since the number of RT items varied from 5 to 10 across the collection of documents, each type of "loss" impacted the RT totals by MT, as well as by wh-type, by different amounts. We effectively lost RT items from one viewing sequence when one of the originally 60 scheduled subjects did not show, and we lost RT items from some translated documents that were viewed but then lost during a network crash.

Wh- MT type	Correct Response Rate	Non-Response Rate	Incorrect Response Rate	Total # RT Items	Total # Subject Responses
When 1	.333	.218	.148	881	696
Where 1	.387	.211	.173	1107	904
Who 1	.417	.120	.154	1103	1159
When 2	.474	.178	.127	875	715
Where 2	.515	.214	.145	1094	920
Who 2	.481	.167	.087	1097	1001
When 3	.410	.216	.173	879	807
Where 3	.443	.207	.173	1103	966
Who 3	.472	.151	.193	1104	1069

Table 6: Event rates tallied over translated-document by subject cases within nine MT by Wh-type categories

strictly increasing with significant differences from When to Where to Who. The non-response rate is clearly lowest for Who, but the three incorrect rates are very close.

The only prediction for subject performance by wh-type that we had before the experiment was that Who items would be easier to detect correctly than When and Where, because the latter could be more complex syntactically and thus more likely to not translate correctly. This indeed occurred, with all three event rates strongest on Who items (highest correct rate, lowest non-response rate, and lowest incorrect rate).

Event Rates by MT Engine and Wh-type

In Table 6, vinally, we tally the MT-by-Wh cross-classified event-rates. Within this Table, the rates evidently vary a great deal across MT-by-Wh categories. In fact, chi-squared tests for "interaction" of log-odds of response between the MT and Wh classifications, with respect to event-rates, show no significant interaction for differences in correct response rates (chi-square 8.96, p-value .061 on 4 degrees of freedom), but highly significant difference in both error rates, the 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).

The presence of interactions between MT and Wh classifications on error rates reinforces what we already observed in the comparative MT engines analyses: that the lack of uniform results pre-empts an across-the-board ranking of the MT engines.

6 Conclusions and Future Work

This paper reports on the results of a one-of-a-kind, large-scale, MT evaluation experiment where nearly sixty subjects extracted who, when, and where-type elements of information from output generated by three types of Arabic-English MT engines: a rule-based engine, a morpho-lexical substitution engine, and a statistically-trained engine. The statistically significant results were mixed, pre-empting an analysis with a single, across-the-board ranking of the engines. To assist users in interpreting these results for their own work environments, we are now working with the collected data to develop a single customizable metric with loss functions, weighted over these rates.

While the experiment and evaluation methodology have provided the results that the MT users request, namely an analysis of a real-world task on multiple MT engines, this approach is quite costly, time-consuming and labor-intensive. As Coughlin (2003) has noted, resource considerations such as these have forced the field to rely heavily on automated metrics. Thus, it is crucial in any evaluation to determine how well results with these metrics compare to the results we find in task-based analyses. In particular, we want to know whether there is a relationship between these popular, strictly text-based metrics and the end-to-end (machine and user) effectiveness metrics of concern to real users.

Our next step is to explore this relationship by studying various aspects of automated metric correlation with subject responses from the information extraction task. This work motivates the need to extend testing methods beyond correlations and to develop other uses of these metrics for a more user-centered evaluation. Thus, we are testing for whether it will be possible to leverage the collected data using it in statistical models that we build and test for their adequacy in predicting task results quickly and less expensively.

	MT1	MT2	MT3
Bleu 4gm	0.0911	0.2347	0.0553
Open-class 1gm	0.4728	0.6128	0.4451

Table 7: Bleu 4-gm and Open-class 1-gm scores

	MT-1 vs MT-2	MT-2 vs MT-3	MT-1 vs MT-3
Bleu 4gm	MT-2	MT-2	MT-1
Open-class 1gm	MT-2	MT-2	Not signif.

Table 8: Wilcoxon Test Results on Scores in Table 7

Toward this goal, we calculated automatic n-gram metrics for output-text translation accuracy, BLEU 4gm and Open-class 1gm metrics, on the machine-translated documents from the experiment (see Table 7). Preliminary results with pairwise Wilcoxon tests (see Table 8) indicate that, for the Bleu 4gram and Open-class 1gm metrics, there is a statistically significant difference across the board with MT-2 scoring decidedly higher than both MT-1 and MT-3. It is only on Bleu 4gm that a statistically significant difference favors MT-1 over MT-3.

While these results yield a ranking of the MT engines that is not inconsistent with some aspects of the task-based results, the dataset from the extraction experiment will support more fine-grained analyses and tests beyond simple rank correlations.[9] The initial testing of a range of generalized linear models indicates that such automatic metrics alone are not sufficient as explanatory variables to yield an adequate model for predicting task response results (Tate 2005). Further model-testing is now underway exploring a broader range of text-based metrics, including source language complexity (Clifford et al., 2004).

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 (Artis-Tech, 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

Banerjee, S. & Lavie, A. 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.

Calison-Burch et al. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the European Association for Computational Linguistics* (EACL-2006). Trento, Italy.

Coughlin, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality. In *Proceedings of MT Summit IX*. New Orleans, LA. pp. 63-70.

Church, K., & Hovy, E. (1993). Good Applications for Crummy Machine Translation. *Machine Translation*,8.

Clifford, R., Granoien, N., Jones, D., Shen, W., & Weinstein,C. (2004). The Effect of Text Difficulty on Machine Translation Performance. In *Proceedings of Language Resources and Evaluation Conference (LREC-2004)*. Lisbon, Portugal.

Fisher, F., Schlesiger, C., Decrozant, L., Zuba, R., Holland, M., & Voss, C.R. (1999). Searching and Translating Arabic Documents on a Mobile Platform. In *Proceedings of the Advanced Information Processing and Analysis Conference* (AIPA-99). McLean, VA.

Holland, R. (2005). Embedded Machine Translation Prototypes at MITRE. Presentation at UMIACS Computational Linguistics Colloquium. University of Maryland, College Park. July 27, 2005.

Laoudi, J., Tate, C., & Voss, C. (2006). Task-based MT Evaluation: From Who/When/Where Extraction to Event Understanding. In *Proceedings of the Language Resources and Evaluation Conference* (LREC-2006). Genoa, Italy.

Lavie, A., Sagae, K., & Jayaraman, J. (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.

Liu, D. & Guildea, D. (2005). Syntactic Features for Evaluation of Machine Translation. In *Proceedings*

[9] Just as the analogy has been made that no single house is suited to all people and no single MT system is suited to all people, it is also true that no MT metric is suited to everyone's needs. While automatic metrics have been used primarily by the MT development community, we are nonetheless interested in how they may be of use to the MT user community. The pace at which new MT metrics are being introduced is invigorating the field and encouraging wider participation and challenges that will ultimately lead to metrics that will be better understood by all. Recent proposals include (i) detecting paraphrased variants of reference translations (Banerjee & Lavie, 2005), (ii) identifying output of syntactic forms that correlate highly with human subjective judgments (Liu & Gildea, 2005), and (iii) capturing more highly valued information in content words with part-of-speech re-weighting (Callison-Burch et al., 2006).

of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, MI.

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

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

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

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

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

Spaerck-Jones, K., & Gallier, J.R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, Berlin.

Tate, C.R. (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.

Tate, C.R. (2003). Task-based MT Evaluation: Tackling Software, Experimental Design, & Statistical Models. Proceedings of Workshop on Machine Translation Evaluation Towards Systematizing MT Evaluation. MT Summit IX. New Orleans, LA.

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

Vanni, M., Voss, C.R., & Tate, C.R. (2004). Ground Truth, Reference Truth & "Omniscient Truth" - Parallel Phrases in Parallel Texts for MT Evaluation. *Proceedings of LREC*. Lisbon, Portugal.

Voss, C.R. (2002). MT Evaluation: Measures of Effectiveness in Document Exploitation. *DARPA TIDES PI Meeting*. Santa Monica, CA.

White, J., 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.