The Naming of Things and the Confusion of Tongues:

An MT Metric

Florence Reeder‡, Keith Miller‡, Jennifer Doyon§, John White§

‡The MITRE Corporation
7515 Colshire Drive, McLean VA 22102
USA
{freeder, keith}@mitre.org

§Litton PRC
1500 PRC Drive, McLean VA 22102
USA
{doyon_jennifer, white_john}@prc.com

Abstract

This paper reports the results of an experiment in machine translation (MT) evaluation, designed to determine whether easily/rapidly collected metrics can predict the human generated quality parameters of MT output. In this experiment we evaluated a system's ability to translate named entities, and compared this measure with previous evaluation scores of fidelity and intelligibility. There are two significant benefits potentially associated with a correlation between traditional MT measures and named entity scores: the ability to automate named entity scoring and thus MT scoring; and insights into the linguistic aspects of task-based uses of MT, as captured in previous studies.

1 Introduction

Machine Translation (MT) Evaluation (MTE) has lacked standards for metrics since the beginning of MT technology, despite concerted efforts throughout its history. There is no meaningful ground truth for MT, since anything can be expressed and translated correctly in many different ways. Statistically valid subjective judgments must be captured, a long and arduous process. Also, it is difficult to associate these quality judgments with the applicability of MT systems to actual usefulness (White, 2000a). At the same time, the vision of instantaneous information access regardless of source language requires very rapid evaluation of approaches and systems. Furthermore, for a view of MT as embedded in other processing, evaluation metrics must be meaningful to the functions to which the output of MT will be applied; replicable for new systems and domains; and automated to the greatest degree possible.

This paper presents a step in the development of a new evaluation approach, which takes advantage of readily measurable phenomena to predict the much harder-to-measure properties of MT output. Section 2 discusses previous and contemporary MTE methods. Section 3 describes the experiment of arriving at named-entity scores from MT system outputs for comparison of different system qualities. After this, it describes how these scores will be compared to existing quality scores. Section 4 presents the results of the first experiment, followed by analysis and future research directions.

2 Previous Evaluations

The 1994 DARPA MT Evaluations (White et al., 1994) were part of a series designed to capture the extensibility of MT approaches to potential applicability within the variety of tasks which could benefit from translated information. Although not the first to suggest the correlation between MT quality needs and user requirements (see Van Slype, 1979), it tackled the issues of: eliciting dimensions of judgments from otherwise disinterested target-native subjects; capturing judgments with finer granularity than before; and using sufficient human factors controls to show reasonably valid measures of fidelity and intelligibility. In this series of evaluations, system outputs were compared to human translations on the criteria of adequacy, informativeness (both measures of fidelity) and fluency (a measure of intelligibility). As described briefly in section 3, the criteria were elicited from subjects whose rating was on a holistic 1-5 scale.

While revealing in the context of the program in which it was administered, these results did not directly serve the larger needs of MT evaluations to give feedback to users, developers and funding agencies. Nor did they meet the desired qualities of meaningfulness, replicability or automation. The DARPA evaluation was expensive and time-consuming because of the need for a) two expert translations of input texts and b) an elaborate test design and administration procedure, with a large number of input documents, output documents, human subjects and decision points for measurement and analysis. Moreover, the DARPA series did not directly provide insight into the place of MT in the continuum of language processing and NLP uses.

From the 1992-1994 series, and the aftermath that followed, a fresh look at MT and MTE arose. As part of this, the notion of examining MT in light of what it is to be used for came to pass. Yet, treating MT as part of an information processing flow does not reduce the amount of work to be done for MTE. From this realization grew the MT Task Proficiency Scale (Doyon & White, 1998; White & Taylor, 1998) which characterizes the tasks that MT output could be used for in an ordered place in a continuum. If one can, for instance, show the

applicability of an MT system to a more demanding task in the scale, then one could safely assume the use of it in a less demanding task.

While potentially informative, the Task Proficiency Scale still suffers the human subject difficulties present in other studies. New work has focused on ways to automate MTE, without relying on time-intensive, human-subject elicitations. These studies (c.f., Jones & Rusk, 2000; Hirschman, et al., 1999; White 2000b) have raised the possibility that automatic measurement of certain attributes or a set of attributes of MT output might be extrapolated to predict measures germane to MT itself, particularly in light of the task-oriented view.

One such attribute, named entities, holds promise as an attractive possible measure. Named entities fall into delimited categories (proper personal names, corporation names, names of geographical locations, dates and monetary amounts, etc). The set of names in translation has some advantages as a metric over a holistic score for the documents as a whole: for one thing, the set has fewer correct translation possibilities. The processing of translating (and/or transliterating) names is neither trivial nor solved (e.g., Knight & Graehl, 1998), but advances in named entity processing have made it a viable candidate for a metric. Additionally, the importance of named entities in other candidate tasks such as information extraction may give insights into the Task Proficiency Scale.

3 Testing Methodology

Based on the results of previous a named entity translation evaluation (Hirschman, et al., 1999), we performed experiments on a larger corpus, the DARPA 1994 corpus. The largest corpus of a series of evaluations[1] includes 100 newspaper articles in each of three source languages: Spanish, French and Japanese. Each source language had two human translations, performed by expert translators, into English. Each source language was then processed by several MT systems in various states of maturity, again with English as the target language. Each translation (system and human) was subject to three separate evaluation criteria:

- **Adequacy** – the presence of correct meaning in target language MT output. A *fidelity* measure.
- **Informativeness** – a reading comprehension-like test on the translated test. Another *fidelity* measure.
- **Fluency** – the degree to which the text is well-formed English. A *intelligibility* measure.

The evaluation materials were designed and executed to meet specific program goals, yet are still in use today (Doyon, et al., 1998). The results of these evaluations are summarized in Table 1.

SOURCE	Adequacy	Fluency	Inform.
EXPERT	94.5	89.2	85.2
PAHO	82.9	53.1	83.5
SYSTRAN	77.1	39.1	82.2
GLOBALINK	77.0	40.8	82.0
PANGLOSS	52.6	19.3	59.5

Table 1: Scores for DARPA 1994 Evaluation

3.1 The Data

This corpus is appealing because it gives us the opportunity for comparing a new metric with human subjective judgments. It therefore facilitates comparison with other methods, allows discovery of correlations between human judgments and automated metrics. The experiment here uses the Spanish → English language document set.[2] For this reason, we have had to exclude the results from the LINGSTAT system as 20 of the articles are not in the current data set.

3.2 The Procedure

For the purposes of this experiment, we hand annotated one of the expert translations, designated as REFERENCE. The named entities were tagged according the MUC named entity guidelines (MUC, 1998). Annotations were done using the Alembic workbench (Figure 3 at the end of the article). A single annotator was used, although annotations were reviewed.[3] We are planning using multiple annotators and checking for inter-annotator agreement, but felt that a reviewed tagging was sufficient for the first experiment.

The test set was then article-aligned. Each tagged article was aligned with the corresponding article from another group in the data set. That is, REFERENCE was aligned with EXPERT, SYSTRAN, etc. Paragraph-level alignment was initially considered, but rejected when it was realized that the two human translations failed to align because of code set conversion problems between different operating systems (Macintosh Code Page versus Microsoft Code Page). For reasons noted later, more detailed alignments are planned.

We utilized the REFERENCE-EXPERT alignment to serve as a baseline for our scoring algorithm and also to indicate the degree of match in human translations. Initially, only exact matches were considered. This yielded a score of less than 80% names matching. By arriving at a score for the REFERENCE-EXPERT pair and analyzing the mismatches, we could identify a set of constraints that can be reasonably relaxed in the scoring algorithm.

Further examination of the mismatches allowed us to incorporate the following relaxations into the scoring algorithm: normalization of numeric entities,

[1] Human Language Technology (HLT) initiative series (White et al, 1994).

[2] Currently available at :
http://issco-www.unige.ch/projects/isle/mteval-april01/

[3] It may be rightfully argued that multiple annotators should be used.

capitalization and diacritic stripping. In particular, the handling of numeric entities was important: "10" should be scored as matching "ten". On the other hand, one of the sources of difficulty in translation for humans appears to be the handling of numbers, as will be discussed in the analysis section.

Capitalization was later ignored because of the individual differences in translating with regard to "title case". "The Good Housewife" should be equivalent to "the Good Housewife". This affected a very small number of instances in the REFERENCE-EXPERT pairing (less than 10). Diacritics were also stripped due to the inconsistency of human translators in preserving them. This resulted in roughly a five percent improvement in the REFERENCE-EXPERT score. It could, and should, be argued that the relaxation of diacritic matching reflects an overall problem in name translation.

Other sources of mismatch not correctly handled in our current implementation are noted in Table 2. Partial matches can occur because of a) word order differences; b) stop word differences; c) titles of people; and d) keyword differences in the named entity, which is especially prevalent in organization names.

SOURCE OF MISMATCH	NUMBER	%
Country Designation	49	13.3
Inconsistent #	14	3.8
Titles	5	1.4
Singular vs. plural	9	2.4
Acronyms	19	5.2
Partial Match	173	47.0
Other[4]	99	26.9
	368	100

Table 2: Sources of Mismatch in Named Entities

3.3 Scoring

The scoring proceeded as follows: each marked name in the reference text was extracted. Then, duplicates were eliminated. This gave a set of named entities in each article. The corresponding article was then checked for the named entities. Only one instance was required for a match, and again duplicates were ignored. This decision reflects the gross alignment and also the fact that human translations more frequently exhibited ellipsis and other language phenomena not normally present in MT system output.[5] We then normalized all scores against the REFERENCE-EXPERT baseline as it was serving as the "gold standard."

4 Results

The results were not as encouraging as we had hoped. The tagging yielded 2646 unique named entities in the 100 REFERENCE articles. REFERENCE-EXPERT only agreed 86.09% of the time, even with all constraint relaxation. Table 3 shows the scores for the systems. The system scores are shown also as normalized for the human-human scores.

Pairing	Count	Score	Normalized
EXPERT	2278	86.09	100.0
PAHO	1972	74.53	86.6
GLOBALINK	1732	65.46	76.0
PANGLOSS	1671	63.15	73.4
SYSTRAN	1626	61.45	71.4

Table 3: Scores for Individual Systems with Normalization of Scores

It should be noted that the constraint relaxations served to aid certain systems at the expense of others. For instance, PANGLOSS did not preserve capitalization with any regularity which means that the relaxation of the constraint greatly improved its score. Whether this is a reasonable constraint relaxation will emerge from the analysis. The same is true of the diacritic stripping phenomena. While one does not want to punish systems unjustly for removing diacritics, more study of the seriousness of the relaxation is warranted. After getting name scores, we then compared the name score with the adequacy measures. Table 4 captures this information.

SYSTEM	Name Score	Adequacy
EXPERT	86.1	94.5
PAHO	74.5	82.9
SYSTRAN	61.5	77.1
GLOBALINK	65.6	77.0
PANGLOSS	63.2	52.6

Table 4: Adequacy Scores compared with Name Scores

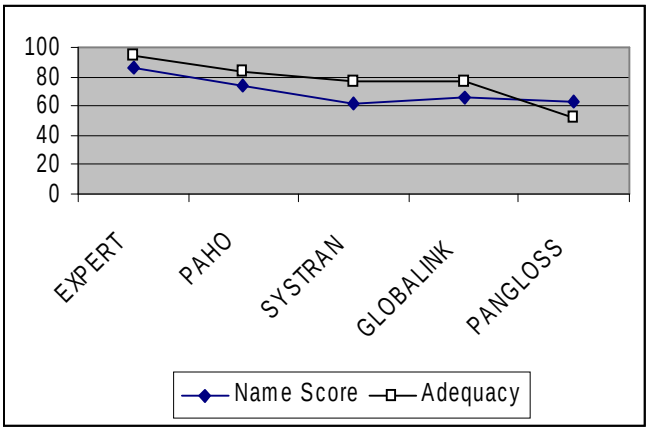

Figure 1: Scores – Adequacy and Name Score

5 Analysis

The question, now, is to look at whether a correlation exists between the named entity scores and other existing DARPA metrics. Figure 1 shows the scores charted at the

[4] Dates / weights / measures / different translations

[5] We recognize that it is these phenomena which could make for a really good translation. It is a subject of future research to determine how closely machine translations should even be compared to human ones.

grosses level of analysis. While there is a rough correlation, based on the average scores for the systems, defined by the relative groupings of the systems by the two scores, it is not a sufficiently strong measure to suggest the clear correlation that we had hoped for. Normalizing the data makes the correlation worse instead of better as shown in Figure 2. We believe that in the case of the PANGLOSS system, relaxing the constraints on capitalization and diacritics gave it a much higher score than anticipated and that normalization benefits it similarly.

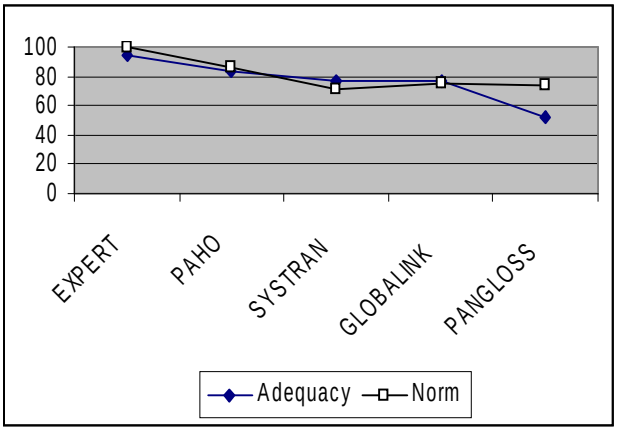

Figure 2: Normalized figures compared to adequacy.

A statistical analysis on an article-by-article basis did not provide support for the hypothesized correlation between these scoring systems with respect to the other systems, using either Spearman rank-difference correlation and the Pearson's correlation coefficient at the 0.05 level. However, this analysis revealed a significant correlation at the article level between the named entity scores and the DARPA adequacy scores in two of the five cases. These initial results may be due to a number of variables which are subject to ongoing analysis, including the variant difficulty levels of the test texts themselves. A more detailed analysis, factoring out additional sources of variation is in order.

There are many possible reasons for this: a) the data analysis is at too rough a grain to be meaningful; b) the article-level scoring skews the contributions of individual named entities; c) the relaxations of the matching criteria definitely favored some systems, particularly those at the lower end of the adequacy scores; d) human scoring may not be as accurate as a less subjective measure; and e) translation success is dependent on more than the sum of the parts of translation. The possibility remains, however, that the named entity evaluation score, while providing interesting and useful predictive information regarding the probable success of various types of downstream processing, may be measuring something different than what is measured by the DARPA adequacy score. Clearly, we have our work cut out for us.

6 Conclusions and Future Research

A correlation between the named-entity scores and the DARPA measures of the same corpus, can imply that using named entity measures will predict the fluency, adequacy, and informativeness attributes of translations. Since no strong correlation exists, we must look to additional scoring techniques to arrive at the overall answer. In fact, building an MTE model may be much like constricting an economic indicator model – many pieces are necessary to capture the true essence of the overall picture. Additional elements to be analyzed include the relationships between named entities, the effects of ellipses and co-reference, and the inclusion of technical and other specialized terminology.

Still, success is ultimately achievable as a correlation between DARPA scores and the MT Task Proficiency scale is possible to establish. Since the translations from the scored DARPA corpus were used in the development of the Task Proficiency Scale, the correlation between these two pieces will contribute to the deeper analysis of the translation process. From these two connections, it will be possible to determine the extent to which named entity translations contribute to the overall success of MT systems in different proficiency tasks.

As noted throughout this work, we still have many questions to be answered, even about this "simple" metric. Our future work focuses on answering them and then extending to other language pairs and corpora. Finally we look to the development of our model of MT system health which combines named entity translation scores with other automated metrics to present a useful picture of MT quality.

Acknowledgements

Our thanks to the anonymous reviewers for their very helpful comments.
Part of this work was performed on the DARPA TIDES program. Approved for Public Release, Distribution unlimited.

Bibliographical References

Doyon, J., Taylor, K., & White, J. 1998. The DARPA Machine Translation Evaluation Methodology: Past and Present. *Proceedings of AMTA-98*. Philadelphia, PA.

Hirschman, L., Reeder, F., Burger, J., & Miller, K. 2000. Name Translation as a Machine Translation Evaluation Task. *Proceedings of the Workshop on Machine Translation Evaluation, LREC-2000*.

Jones, D. & Rusk G.2000. Toward a Scoring Function for Quality-Driven Machine Translation. Proceedings of Coling-2000.

Knight, K. & J. Graehl. 1998. Machine Transliteration. *Computational Linguistics, 24(4)*, pages 598–612.

MUC-7. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7).* http://www.muc.saic.com/proceedings/muc_7_toc.html

Taylor, K. B. & J. S. White 1998. Predicting what MT is Good for: User Judgments and Task Performance. *Proceedings of Third Conference of the Association for Machine Translation in the Americas, AMTA98*. Philadelphia, PA.

White, J. S. 2000b. Toward an Automated, Task-Based MT Evaluation Strategy. Athens, Greece: Proceedings

of the Workshop on Evaluation, Language Resources and Evaluation Conference.

White, J.S. 2000a. Contemplating Automatic MT Evaluation. In J. White (ed.) *Envisioning Machine Translation in the Information Future.* Lecture Notes in Artificial Intelligence #1934. Berlin: Springer.

White, J. S., T. A. O'Connell, & F. E. O'Mara. 1994. Advanced Research Projects Agency Machine Translation Program: 3Q94. Proceedings of the November 1994 Meeting.

REFERENCE TEXT – ANNOTATED

<ENAMEX TYPE="PERSON"> UN Secretary General Boutros Boutros-Ghali </ENAMEX> said <TIMEX TYPE="DATE"> Wednesday </TIMEX> that if international sanctions against <ENAMEX TYPE="LOCATION"> Haiti </ENAMEX> to restore democracy there continue to be ineffective he would favor a "more completely <ENAMEX TYPE="LOCATION"> Haitian </ENAMEX>" solution involving a "compromise" negotiated among the parties. While the <ENAMEX TYPE="LOCATION"> United States </ENAMEX> has recently increased threats of military intervention to force coup members from power, <ENAMEX TYPE="PERSON"> Boutros Ghali </ENAMEX>, in a report to the <ENAMEX TYPE="ORGANIZATION"> UN General Assembly </ENAMEX> <TIMEX TYPE="DATE"> Wednesday </TIMEX>, said he believed a solution to the <ENAMEX TYPE="LOCATION"> Haitian </ENAMEX> crisis could only be obtained "through a compromise based on constructive concessions to which both parties consent."

REFERENCE TEXT

UN Secretary General Boutros Boutros-Ghali said Wednesday that if international sanctions against Haiti to restore democracy there continue to be ineffective he would favor a "more completely Haitian" solution involving a "compromise" negotiated among the parties. While the United States has recently increased threats of military intervention to force coup members from power, Boutros Ghali, in a report to the UN General Assembly Wednesday, said he believed a solution to the Haitian crisis could only be obtained "through a compromise based on constructive concessions to which both parties consent."

SYSTRAN TRANSLATION

The Secretary General of the UN, butros butros-Ghali , "more properly" pronounced the Wednesday in favor of a Haitian solution resultant of a "negotiated commitment" between the parts, if the international sanctions against Haiti continue being ineffective to recover the democracy in that country.

While the United States multiplied the last days the threats of a military participation to force the coup participants to abandon the power, butros ghali considered in a report directed the Wednesday to the general Assembly of the UN that a solution of the Haitian crisis only could be obtained "with a commitment, based on constructive and allowed concessions" by the parts.

Figure 3: Example documents: tagged document, reference text, system output