Edit Distance: A Metric for Machine Translation Evaluation

Mark Przybocki, Gregory Sanders, Audrey Le
National Institute of Standards and Technology (NIST) – Information Access Division, Speech Group
100 Bureau Drive, Stop 8940, Gaithersburg, MD 20899-8940
{Mark.Przybocki, Gregory.Sanders, Audrey.Le}@NIST.gov

Abstract

NIST has coordinated machine translation (MT) evaluations for several years using an automatic and repeatable evaluation measure. Under the Global Autonomous Language Exploitation (GALE) program, NIST is tasked with implementing an edit-distance-based evaluation of MT. Here "edit distance" is defined to be the number of modifications a human editor is required to make to a system translation such that the resulting edited translation contains the complete meaning in easily understandable English, as a single high-quality human reference translation. In preparation for this change in evaluation paradigm, NIST conducted two proof-of-concept exercises specifically designed to probe the data space, to answer questions related to editor agreement, and to establish protocols for the formal GALE evaluations. We report here our experimental design, the data used, and our findings for these exercises.

1. Introduction

Edit distance, as a Machine Translation (MT) metric, is an intuitive measure of the rate of errors in MT output, (number of errors, divided by number of reference words), with each edit viewed as fixing an error. Its use in the DARPA GALE program is motivated by background studies showing what level of edit distance corresponds to MT output that is directly usable by military personnel. In this paper, we report experiments that we have done toward implementation of this metric on all the MT outputs that will be produced in the GALE program. The idea is to measure the minimum number of edits that are needed to make the MT output have exactly the same meaning as, and to be as understandable as, a careful human reference translation. Edit distance is, however, just one metric in a long history of the metrics that have been explored for MT research.

The most widely accepted benchmark for the quality of MT outputs is expert human judgments of their semantic accuracy and fluency (King, 1996). Semantic adequacy judgements may come from either bilingual or monolingual judges. In practice, the human judgments of adequacy and fluency turn out to be somewhat subjective and may be insufficiently fine-grained to measure progress. They, further, are summative rather than formative metrics (Nielsen, 1993), generating little usable information about what is deficient in the MT output.

Semantic adequacy judgments made by bilingual judges such as expert human translators (an approach currently being used in the DARPA TransTAC program) can be viewed as *ground truth*—an answer that you want candidate metrics to yield. In practice, human judgments of semantic adequacy often come from monolingual human judges comparing the MT output to one or more careful human translations. Our implementation of edit distance is based on monolingual editors, which we view as a parallel to that model of monolingual human judges.

The idea of using some measure of edit distance as a metric for MT quality has been explored since at least the early 1990s (see Frederking and Nirenburg, 1994; Knight and Chander, 1994; King, 1996). Use of Word Error Rate combined with sentence-by-sentence choice from multiple reference translations was explored by Niessen *et al.* (2000). The evaluation of MT technology received a boost in 2001 when IBM introduced "BLEU" (Papineni, *et al.*, 2001, 2002), an automatic and repeatable n-gram evaluation metric that demonstrated high correlation with human judgements of system adequacy and fluency (Doddington, 2002). BLEU was used as the primary evaluation metric for the NIST 2002-2005 MT evaluations conducted under DARPA's Translingual Information, Detection, Extraction, and Summarization program. BLEU is a measure of precision. At a later point, Lavie and Banerjee (Banerjee and Lavie, 2004; Lavie, *et al.*, 2005) introduced a somewhat related metric, METEOR, which measures both precision and recall, and finds its set of co-occurrence matches by beginning with exact matches and then extending the set of matches first by stemming and then by matching based on WordNet classes (see http://www.cs.cmu.edu/~alavie/METEOR/).

While acknowledging the value that BLEU has provided to the MT research community in both evaluation and development, it is universally recognized that BLEU (or any other automatic metric for that matter) has not been proven to be effective in assessing the quality of system translations, or in showing how useful the translations might be to an end user or to downstream processing. In an effort to address these issues, NIST will be evaluating MT quality for DARPA's GALE program in terms of edit distance.

Unlike previous NIST MT evaluations that used the BLEU metric, the GALE evaluations using edit distance will be neither automatic nor completely repeatable, since edit distance relies on the decisions of human editors as defined in section 2.

To prepare for the change in evaluation paradigm, the NIST Speech group conducted two proof-of-concept (POC) exercises which were designed to help establish the protocols to be followed for the GALE evaluation. The two exercises where designed around the different data types planned for use in the GALE program. Incremental changes (improvements to the exercise

instructions and to the editing tool itself) were incorporated at the end of each exercise. The two exercises can be summarized by the type of data that was translated:

- POC-1: Post editing using Arabic newswire text, and
- POC-2: Post editing using Chinese newswire text.

The two POC's culminated in a dry run evaluation for the three GALE teams (GALE-DryRun, 2006). In this paper, we will compare the results of the edit distance metric with the results of BLEU, METEOR, and human judgments on our two proof-of-concept exercises.

2. Defining Edit Distance

We define edit distance to be the number of insertions, deletions, and substitutions that are required in order to make a system translation equivalent in meaning to that of a reference translation, using understandable English. In our calculation of edit distance the insertion of two consecutive words counts as two edits. Likewise the deletion of two consecutive words counts as two edits. But moving one or more consecutive words to somewhere else in the translation (shifting a string of any number of words, by any distance) only counts as one edit.

We use publicly available software developed by Snover (Snover, 2005) to calculate edit distance. Translation Error Rate (TER) reports the ratio of the number of edits incurred to the total number of words in the reference text.

An example of calculating edit distance follows.

Original text:

```
To bring an end to military conflict
on October 6 on a a comprehensive
blockade against Palestine .
```

Edited:

```
To bring an end to military conflict ,
the Israeli military began a
comprehensive blockade against
Palestine on October 6 .
```

The example above shows four insertions (added ", the Israeli military"), one substitution (substituted "began" for the second "on"), one deletion (deleted the extra "a"), (moved "on October 6" to the end) for a total of seven edits or errors. The high-quality human reference translation (not shown) for this segment has twenty words. TER for this segment is 7/20 or 35%.

3. Post-editing Guidelines

NIST provided the editors with guidelines that they were to follow while making their edits. The guidelines were updated after each POC exercise based on our review of their edits, and the comments that each editor supplied after completing the task.

Our goal for the guidelines is to develop rules that would promote inter-editor agreement. We did not want the guidelines to be too cumbersome to follow. We didn't expect editors to memorize the entire document, but rather have a reference manual to consult as needed. However, we did supply a one-page high-level set of rules that the editors were to keep by their side until they memorized the most important set of rules. These high-level rules for POC-1 were:

Make the MT output have the *correct meaning*, be *readily understandable*, and really be *English*.

(1) Make the MT output have exactly the same meaning as the reference human translation.
(2) Make the MT output be as understandable as the reference.
(3) Punctuation must be understandable, and the sentence-like units must have sentence-ending punctuation. But do not otherwise insert, delete, or change punctuation merely to follow traditional optional rules about what is "proper".
(4) If words/phrases/punctuation in the MT output or in the reference human translation are completely acceptable, use them rather than inserting or substituting something new and different. Ignore this guideline when it conflicts with the other guidelines – or when words/phrases from the reference need to be inserted and you think up a substantially shorter replacement with exactly the same meaning.
(5) Dates, as well as commas and decimal points in numbers, should be formatted according to U.S. conventions (for example, convert 23-2-2004 to 2-23-2004).
(6) Make the MT output sufficiently fluent that it is really English.

For POC-2, the fourth rule was simplified by removing the last sentence. (*These rules were further simplified before embarking on the GALE dry run evaluation.*)

The guidelines where again updated before the GALE dry run to address characteristics specific to the editing of speech data. The guidelines contain plenty of examples, and the current version can be accessed from the NIST GALE web-site, URL: http://www.nist.gov/speech/tests/GALE/2006dr/doc/GALEpostedit_guidelines-2.0.4.pdf.

4. Post-editing Interface

NIST developed a JAVA based editing tool that is tailored to the task of post editing. Editors were instructed to consider the document as a whole but were only able to edit one segment at a time. To assist the editor, the complete context of the document was provided in three rows. The first row contains everything already seen in the document, the second row contains the active editing segment, and the third row contains all the remaining segments. The reference data is in the first column and the aligned translation to be edited is in the second column. The tool contains a third column designed to provide edit distance feedback to the editor.

The upper rightmost cell uses the "diff" function to show the difference between the original MT and the edited-thus-far version. It is a rudimentary count of the edit distance. The middle cell in the 3rd column always

contains the original MT output. This gives the editor a reference point if they ever want to re-edit a segment. The bottom cell uses the "diff" function to show the difference between the edited MT and the gold-standard reference. The entire third column may be suppressed at any time.

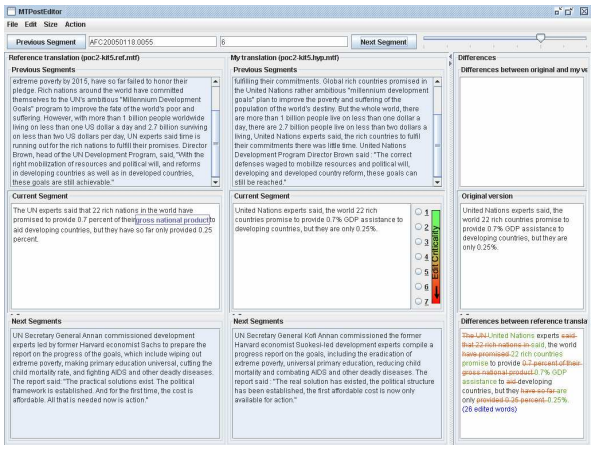

Figure 1: Screenshot of the NIST post editing tool.

5. Proof-of-concept Exercises

The two separate proof-of-concept (POC) exercises were designed to probe the use of the edit distance metric using the types of data to be used in the GALE program. Each POC exercise involved the same cycle: creating the data pool, having editors edit the data, analyzing the resulting edits, and implementing any lessons learned and editor feedback into the next POC exercise.

5.1 POC-1 Arabic Text Data

In planning stages, it was believed that editing Arabic MT would be a slightly easier task than editing Chinese MT or speech MT. This assumption was based on the NIST MT evaluation results that showed higher BLEU scores for the Arabic test set than that of the Chinese text data.

5.1.1 POC-1 Data

NIST had at its disposal a variety of MT translations from the past NIST MT evaluations. For the first exercise we selected 10 Arabic newswire text documents that were used in the NIST 2004 MT evaluation (MT-04). The documents were selected to represent a varied level of difficulty. This was accomplished by establishing an average BLEU score for each document in the MT-04 test set (consisting of 100 newswire documents) and selecting the 5^{th}, 15^{th}, 25^{th}, 35^{th}, …, 95^{th} document when ranked by BLEU score.

NIST selected the translation from three evaluation participants. The systems were chosen to account for varied system performance on the entire Arabic test set, as determined by their BLEU score (BLEU scores range in value from 0.00 to 1.00). The three systems chosen had BLEU scores of 0.31, 0.20, and 0.17 over the set of 10 documents, which ranked them in the top, middle, and bottom scoring groups respectively.

The NIST MT-04 evaluation used four independent translations of high quality for reference. Although each of these translations were considered to be of "high quality", in order to mimic what was envisioned for the GALE evaluations, NIST adjudicated the four references into a single gold-standard reference. We enlisted the help of a native Arabic speaker to view the source documents in cases of ambiguity or to resolve the differences between two or more references.

There were a total of 81 segments (sentence like units) and 2080 words in the 10-document reference set.

Three editing kits were created with each kit containing the 10 documents from the three systems. The order of the documents in each kit was randomly generated.

5.1.2 POC-1 Editors

We recruited volunteers who found the task of post editing to be interesting and who could commit to completing the task in full, in our allotted time. Some of our volunteers were NIST scientists who may not possess the specific skills we will require for our post editing task in the real evaluations. Other volunteers were closer to what we were identifying as possible candidates as editors for the actual evaluation.

For the GALE evaluation we will require that post editors possess the following qualifications:

1. Have reasonable proficiency at using graphical user interfaces on a computer
2. Be a native speaker of English
3. Have a very high level of proficiency at reading and writing English, including ability at editing written English.

For POC-1 we collected edits from 5 volunteers, four of which were NIST scientists, and 1 who was a high-school English teacher. Each editor edited all 10 documents for each of the three systems.

5.1.3 POC-1 Exercise Findings

Our initial goal for the first proof-of-concept exercise was to test our post editing concepts and to use what we learn to develop evaluation protocols.

Table 1 shows some common statistics for our five editors for each of the three systems.

	Sys-1	***Sys-2***	***Sys-3***
#sys words	2020	2200	2520
High edits	649	1052	1178
Low edits	584	891	962
Mean edits	609.2	982.2	1083
Std. Dev.	25.2	68.2	88.2
Variance	633.2	4650.7	7773.5

Table 1: POC-1 Editing statistics using five post editors

Figure 2 shows the total edits each editor made for each set of 10 documents, for each of the three sets of system translations.

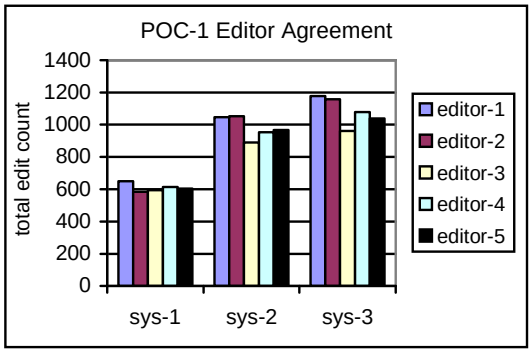

Figure 2: POC-1 Post Editor Agreement

"Sys-1" was the system that received the highest overall BLEU score and for this system we note the overall high editor agreement. The "noisier" the data, the more room for clever editing, thus the growing disparity in total number of edits for system 2 and 3.

The average rate of editing for our five post editors was: 530, 750, 785, 810, and 830 words per hour. Our editor working at the slowest pace found the task to be very burdensome; the other four found the task to be interesting and may be a better estimate of the expected rate of editing.

	Editor 1	Editor 2	Editor 3	Editor 4	Editor 5
Sys-1	27.93	25.13	25.52	26.46	26.03
Sys-2	45.09	45.27	38.34	41.44	41.61
Sys-3	50.69	49.78	41.39	46.43	44.71

Table 2: POC-1 TER scores for each editor

We see from Table 2 that the better the original MT, the more consistent the post editors were in determining how many edits were needed.

5.2 POC-2: Chinese Text Data

The second POC exercise made use of Chinese newswire texts from the NIST 2005 MT evaluation (MT-05). The exercise was designed to expose any nuances of the Chinese language that would pose problems not discovered during POC-1.

5.2.1 POC-2 Data

While POC-1 used system translations from MT-04, MT-05 contained translations from significantly better systems for Chinese, and it would be more appropriate to use such data for a realistic post editing task.

For this exercise we selected 25 Chinese newswire documents. These documents were selected in a similar fashion as those for POC-1 but for every 4th document from a ranked order of average BLEU scores (averaged over the top 7 MT-05 evaluation participants). There were a total of 272 segments (sentence-like units) and 7605 words in the 25 document reference set.

For POC-2 we chose to use the translations from the two top performing systems in MT-05. These systems had a BLEU score of 0.22 and 0.20 for the selected 25 documents. The two translations of each of the 25 documents, gave us 50 translated documents in this experiment. We did a slow, careful, painstakingly systematic human judgment of semantic adequacy for each translated segment in each of these 50 documents.

Unlike POC-1 where each editor edited all documents from all three systems, it was not feasible to have each editor editing 50 documents. In an effort to reduce editor fatigue (one of the comments from our editors in POC-1) we decided to have each editor edit only 10 documents.

Therefore, we divided the 50 documents into five datasets with each dataset having only 10 documents. Each dataset had an equivalent distribution of BLEU scores, so the five datasets were presumed to be of equal difficulty. Each dataset was made up into two editing kits, with the ten documents in the opposite order of presentation and with documents from the two systems alternating. Each of the 50 documents was therefore edited by a minimum of two editors.

5.2.2 POC-2 Editors

Again, we relied on volunteers to perform the post editing. For POC-2 we recruited 12 volunteers all with varied backgrounds of editing experience. Two of our editors from POC-1 participated in POC-2 allowing for some direct comparisons.

Regrettably, the two editors for one of the datasets (*dateset 4*) were not able to complete the task as requested; we were able to determine that one editor simply did not have sufficient available time to complete the editing. We have excluded that dataset from the results presented here.

5.2.3 POC-2 Exercise Findings

Our analysis of the results of POC-2 focused on comparing the edit distance results to the results of the automated metrics (BLEU and METEOR) and to the human judgments of semantic accuracy and fluency. The edit distance results are shown in Figure 3 and Figure 4, below. Each bar in these charts represents a different editor.

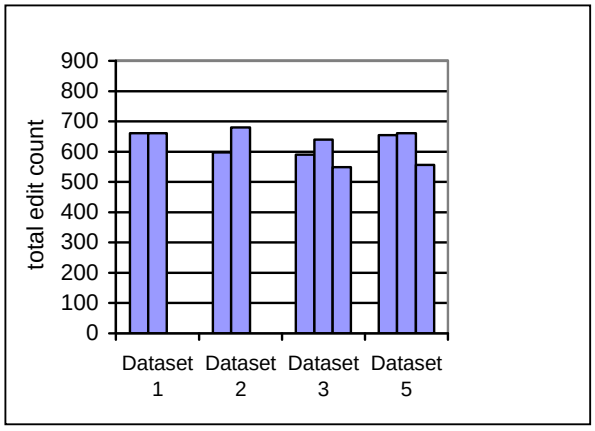

Figure 3: POC-2 post-editor agreement for System-1

As can be seen by comparing Figure 3 with Figure 4, using BLEU scores to choose datasets of equal difficulty (predicting equal edit distance) was more effective in the case of System 1 (above, where the clusters are of fairly equal height) than was the case for System 2 (below, where the cluster for dataset 3 appears better than the others). One can also see that there were fewer edits for System 1 than for System 2: an edit distance result that reflects the better performance of System 1 as measured by BLEU.

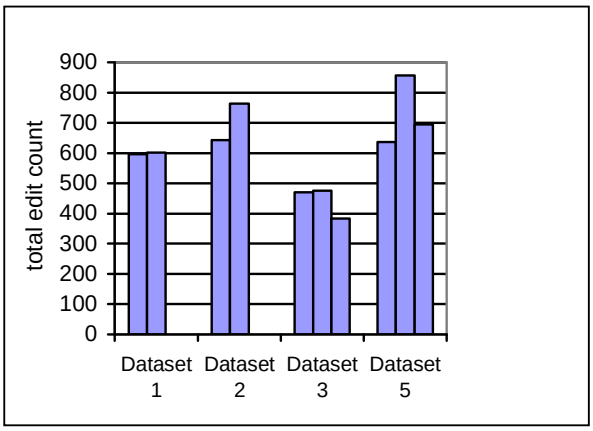

Figure 4: POC-2 post-editor agreement for System-2

As mentioned in the introduction of this paper, careful human judgments of semantic adequacy are sometimes regarded as the gold standard measure of MT quality. We turn now to examining correlations between edit distance and our careful human judgments of semantic adequacy. We also examine correlations of our human judgments with the two automated metrics (BLEU and METEOR) to which we are comparing edit distance. Because each editor edited ten documents, we can look at the correlation for each editor. Further, we regard the correlation values for the editors as independent of each other. Treating the human judgments as a gold standard benchmark, the following table shows the distribution of those correlation values, in order to compare the performance of edit distance to the performance of the two automatic metrics.

	Mean	StdDev	StdErr
Edit distance	0.831	0.087	0.027
BLEU	0.764	0.089	0.028
METEOR	0.789	0.078	0.025

Table 3: Mean Pearson correlation with human judgments of semantic adequacy, across editors

As can be seen by the data for our POC-2 exercise, edit distance correlated (across editors) with the human judgments more strongly (0.831) than did BLEU (0.764) or METEOR (0.789).

Further, this difference in correlation has the suggestion of significance to it. If we compare the mean correlation (across editors) for edit distance to the mean correlation (across editors) for BLEU, the difference in the mean correlations (0.0831 – 0.764 = 0.067) is 2.45 times the standard error of the mean correlations for edit distance. Similarly, for METEOR the difference is 1.71 times the standard error of the mean correlations for edit distance. METEOR thus far seems to have higher correlation with human judgments than does BLEU, probably because it also considers recall.

6. Summary

In this paper we have discussed the two proof-of-concept exercises that were completed in preparation for the GALE evaluations. We found that it is possible to achieve similar editing costs among several editors for automatically produced translations, and the better the system translations, the closer to agreement the editors come.

We found that our definition of the edit distance metric correlates with human judgements of adequacy, as well as or even better than some well accepted automatic metrics.

The exercises provided insight into how to best implement a post editing-based evaluation. Issues concerning editor fatigue were found and alternative plans for editor workloads can be made. Calculations of the rate of editing will allow for better planning of the time needed for evaluation.

A third POC was planned to address “transcription” which for GALE is defined as going from foreign audio to English text. In a short period of time we managed to have one audio broadcast of a “call-in talk show” edited by four editors. Although we did not obtain much data to analyze, the comments by the editors were all very telling of the difficulty of the task.

7. Future Work

Our POC exercises provided us with insight into what we might expect in a formal evaluation of machine translation when using edit distance as the metric. These exercises helped define the protocols that were used in the GALE Translation dry run (GALE-DryRun, 2006) The first formal GALE Translation evaluation will occur in the June/July 2006 timeframe with post editing of the translations occurring during August.

While our POC exercises focused on the ability to achieve acceptable inter-editor agreement when editing for the correct meaning and for fluency, the GALE program will de-emphasize fluency while adding a much stronger emphasis on minimizing the number of edits in order to find a limiting minimum value for edit distance.

8. References

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology* conference notebook, pp. 128–132.

Frederking, R. and S. Nirenburg (1994). Three heads are better than one. In *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP)*, pp. 95–100.

GALE-DryRun (2006). Winter 2006 (GALE-06W) GALE Dry-Run Evaluation Plan. http://www.nist.gov/speech/tests/gale/2006dr/doc/GALE06_dryrun_evalplan.v5.pdf

King, M. (1996). Evaluating natural language processing systems. In *Communications of the ACM* (39)1, pp. 73–79.

Knight, K. and I. Chander (1994). Automated postediting of documents. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Lavie, A., K. Sagae, and S. 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)*. Washington, DC.

Banerjee, S., and A. Lavie (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinisic and Extrinsic Evaluation Measures for MT and/or Summarization* at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). Ann Arbor, MI.

Nielsen, J. (1993). *Usability engineering*. New York: Academic Press.

Niessen, S., F. J. Och, G. Leusch, and H. Ney (2000). An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pp. 39–45. Athens, Greece.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2001). BLEU: A method for automatic evaluation of machine translation. (An IBM research report available at http://domino.watson.ibm.com/library/cybergig.nsf/).

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*: pp. 311–318.

Snover M., B. Dorr, R. Schwartz, J. Makhoul, L. Micciulla, and R. Weischedel. *A Study of Translation Error Rate with Targetted Human Annotation*. LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park, MD, July, 2005.