Paralinguist Assessment Decision Factors For Machine Translation Output: A Case Study

Carol Van Ess-Dykema, Jocelyn Phillips
National Virtual Translation Center
Washington, DC 20535

Carol.j.vaness-dykema@ugov.gov
Jocelyin.h.phillips@ugov.gov

Florence Reeder, Laurie Gerber
MITRE Corporation
7515 Colshire Dr.
McLean, VA 22102
Florence.m.reeder@ugov.gov
Laurie.M.Gerber@ugov.gov

Abstract

We describe a case study that presents a framework for examining whether Machine Translation (MT) output enables translation professionals to translate faster while at the same time producing better quality translations than without MT output. We seek to find decision factors that enable a translation professional, known as a Paralinguist, to determine whether MT output is of sufficient quality to serve as a "seed translation" for post-editors. The decision factors, unlike MT developers' automatic metrics, must function without a reference translation. We also examine the correlation of MT developers' automatic metrics with error annotators' assessments of post-edited translations.

1 Introduction

Machine Translation (MT) is an application not yet in wide use among translation enterprises. MT can perform translations without human intervention or provide "seed translations" for human post-editing[1]. These "seed translations" have the promise of improving translation consistency and speed.

With respect to its potential in a human translation environment, two questions about MT emerge. The first is whether a particular MT system is sufficiently advanced to provide human post-editable "seed translations," and whether the technology constitutes a significant step forward for actual operational use, represented by time savings and by quality improvements.

The second question is whether we can determine which of the available MT output measures best predicts successful human post-editing of MT output. This prediction has two aspects:

- predicting whether a given MT system is of sufficient maturity to be used in the workflow to produce "seed translations;" and
- predicting whether an individual document can be successfully post-edited.

We address both of these in this paper.

Up until now, much of the assessment of MT quality has been focused on supporting the developers of MT systems rather than the users of MT output. Thus it remains to be seen if any of the MT developers' metrics can serve as decision factors for choosing to post-edit a "seed translation" rather than translate a document from scratch. It is difficult to envision a scenario in which any MT developers' metrics would be of use, since MT developers' metrics tend to require *reference translations* (high-quality human translations of a given document).

We will explore what measures of MT output could be used on individual documents if MT developers' metrics are not useful decision factors.

[1] NVTC thanks the Pan-American Health Organization for informing our understanding of the role of the post-editor in the translation process.

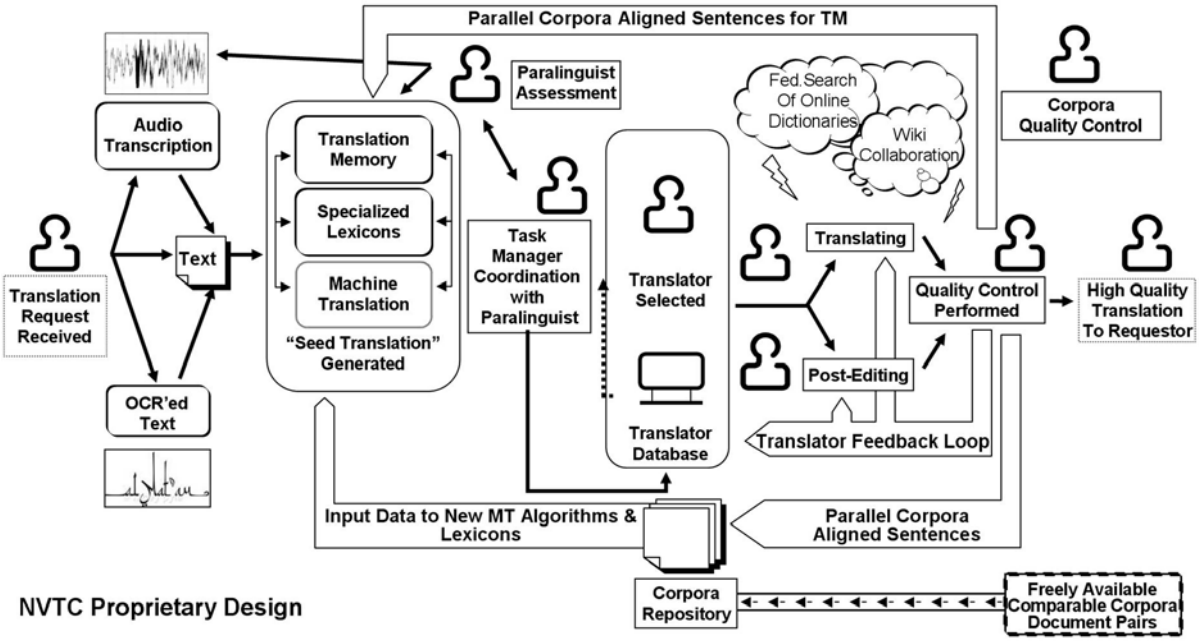

Figure 1: NVTC New Workflow

This paper describes a methodology under development which addresses the two questions as hypotheses. We first present the workflow that the National Virtual Translation Center (NVTC) is currently adopting. We describe the error scoring methodology and the MT developers' metrics used in the investigation. Next, we present MT output metrics that do not use reference translations. Finally, we give the specifics of our on-going case study.

2 Background

The study takes place at the NVTC. The NVTC provides high-quality translations for the U. S. Department of Defense and the Intelligence Community. The Center has translated in well over 100 languages, in many critical subject domains and in over more than 20 genres.

NVTC translators work both onsite and offline, connected virtually to a translation management system, and eventually to a shared tool space. The NVTC's ability to surge and shift to meet emerging needs in new domains and languages makes it unique in U. S. government translating, yet it also lends a number of challenges to technology and process. In particular, a partial dependence on external translators results in idiosyncratic, personal translation tools and processes, which, though flexible and rapidly responsive, result in a loss in consistency and a failure to gain domain and genre knowledge enterprise-wide.

At the same time, world situations change very quickly, and an agile response to new intelligence needs involves the rapid ramping up of translation coverage for additional foreign languages. MT holds promise in this area. However, the extent of improvement to the human translator accuracy, fluency and speed, especially given a relatively low availability of MT systems in some of these languages, remains to be seen. In light of these challenges, NVTC has proposed a translation workflow to meet its immediate and ever-changing translation requirements. Based on the findings of this initial study, NVTC intends to incorporate MT as one of the components in its new workflow. Consequently, the study questions are germane to the workflow. A diagram of the NVTC envisioned workflow for utilizing MT technology is given in Figure 1.

Of note in the workflow is the role of the Paralinguist[2]. The paralinguist identifies the optimal selection of specialized lexicons, translation memory, and machine translation to produce a "seed translation." The paralinguist analyzes the technology-produced output at the document level and directs the document to a human translator to translate "from scratch," or to a human post-editor

[2] NVTC thanks the Canadian [National] Translation Bureau for introducing us to the term "paralinguist", described to us as analogous to the term and function of a paralegal professional.

to convert the output "seed translation" into a finished product. The paralinguist uses values calculated for decision factors to determine the best document routing. These decision factor values are essential because the paralinguist may know only one of the languages being translated. Additionally, the paralinguist may use decision factor values to characterize the kind of translator / post-editor that the text should be assigned to. This assignment will take into account factors such as the genre and domain of the document.

For the purposes of this study, the translation professional's workflow role is one of post-editing MT output. Translator post-editing of MT output transforms it into high quality translations.

Working in a research role rather than a workflow role, the error annotator assesses the document for correctness and completeness once it has been translated from scratch or post-edited. The error annotator will use a research scoring methodology that we describe in Section 3.1

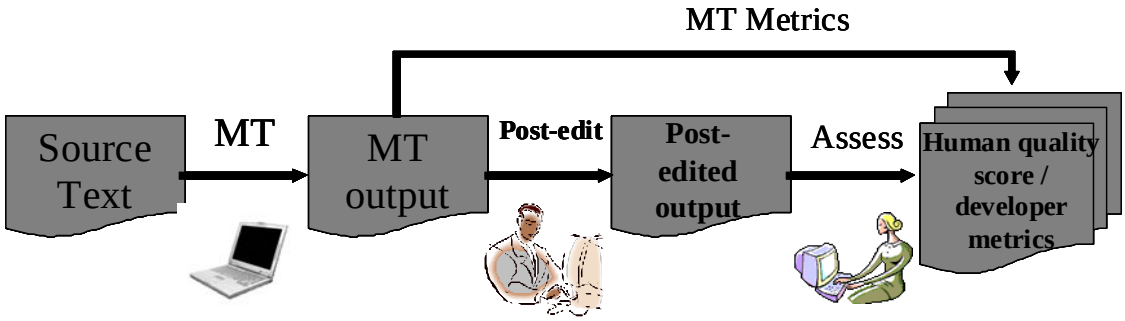

Figure 2: Post-editing of "Seed Translations"

3 Investigation 1

The case study consists of two investigations. The first investigation (Figure 2) answers the questions:

- Can we post-edit MT produced "seed translations" while increasing translation speed and accuracy? We hypothesize that MT helps translation professionals work faster, producing the same quality or higher quality translation.
- Do the translators' opinions of MT output correlate with their speed and accuracy performance?
- Which of the MT developers' metrics predicts successful human post-editing at the system level?

We have designed a series of steps to answer these questions.

1. We use machine translation to translate candidate texts, selected on the basis of subject and genre.
2. Translation professionals are asked to post-edit the output of the MT and we measure their "words per hour translation rate" as the test condition. In the control case, we ask translation professionals to translate the texts directly.
3. Error annotators assess the post-edited MT output using a US Government Translation Proficiency Metric described in Section 3.1.1.
4. Analysis consists of comparing translator speed and accuracy for test and control conditions.

If the post-edited output is of the same or better quality and the time taken to post-edit is less in the test case, then the outcome supports an affirmative answer to the question of whether MT is sufficiently advanced to provide human post-editable "seed translations."

Additionally, we compare MT developers' metrics with translator words per hour; translators' opinion of the post-editing activity and the error annotators' scores to determine whether MT developers' metrics of the raw MT output can be used as predictors of successful human post-editing at the system level.

The design of these investigations is suitable to point to potential trends associating particular metrics with particular post-editing outcomes. This "trend" approach enables the resolution of promising trends without requiring the sample size necessary for full, statistically significant test results.

Once the data have been collected, we will continue the case study with a simulated workflow validation. In a simulated workflow validation, we will provide the paralinguist with documents and their decision factors scores. We then ask the paralinguist to route the "seed translation" based on the decision factors. In this way we expect to show that it is possible for the paralinguist to make these determinations with the decision factor values presented.

3.1 Research Scoring Methodology

In order to evaluate the quality of the post-edited and scratch translations, we considered two scoring methodologies. The first one is commonly used within the US Government to evaluate human translation. The second is the SAE-J2450 (SAE, 2005) translation quality metric. We describe each of these below.

3.1.1 US Government Translation Proficiency Metric

The US Government metric, hereafter referred to as the translator proficiency metric, was developed to assess human translation proficiency. It too has been adopted and adapted for use in other situations, such as machine translation evaluation (Reeder, 2006). The metric consists of two elements:

1. Five error categories;
2. Three meta-rules for multi-error situations.

A human scorer uses the guidelines provided to rate the quality of the given translation. Errors are assigned to one of five error categories. They are linguistically motivated and include the following:

1. Syntactic error;
2. Lexical error;
3. Omission error;
4. Awkward usage error;
5. Punctuation error.

When scoring human translations, the scorers assign a numeric score to each error based on the error category. These scores are then subtracted from the number of words with a threshold determining pass/fail. A score that drops below 5% of the word count is considered a failing grade. In this scoring scheme, a higher score is preferable. Since we are not considering whether an individual has passed a test, but instead are considering the type and number of errors, we have adapted the metric slightly. Our score is not a single number, but instead is a collection containing the counts for each category. In our adaptation, lower scores are preferable.

3.1.2 The Society of Automotive Engineers (SAE) Metric

The SAE metric was developed to standardize the measurement of quality for translation in the automotive industry, regardless of whether the translation is performed by humans, assisted by computer, or performed by computer. As an industry standard, it has been both adopted and adapted for use in other industries (e.g., Schütz, 1999). While the translation proficiency metric contains only two elements, this metric consists of four elements: seven error categories; a severity indicator (minor or severe); two meta-rules for multi-error situations; and numeric weights for error category and severity.

A human evaluator uses the guidelines provided to rate the quality of the given translation. Each error is marked, given a severity indicator and assigned to an error category. The linguistically motivated error categories include the following: wrong term (lexical); wrong meaning (lexical); omission; structural error (syntactic); misspelling error (punctuation); punctuation error; and miscellaneous error.

Each category and severity combination has a numeric score assigned to it. The numeric scores are then aggregated and conditioned by the number of words in the document. This yields a final score where lower scores are preferable.

3.1.3 Selection of Scoring Methodology

Both evaluation metrics have merit, but we selected the translation proficiency metric because it has fewer categories and we therefore believe it is easier for the error annotators to learn and faster for them to apply. Additionally, we are using the translation proficiency metric for this case study because we are currently using it in a translation memory study with the National Institute of Standards and Technology (NIST) and the Naval Research Laboratory (NRL).

3.2 MT Developers' Metrics

We use MT developers' metrics to assess the quality of the MT output. The number of MT developers' metrics has exploded over recent years, spurred on by the automatic metric, BLEU (Papineni, et al., 2001). Included in this list are METEOR (Banerjee & Lavie, 2005), N-gram metric (Doddington, 2002), and BLANC (Lita et al., 2005). These metrics are readily available. We intend to use them to see whether there is a correlation between overall MT system performance and the performance of the human translators in the post-editing task. If the metrics scores correlate with post-editing scores, we may be able to use the metrics to assess readiness of systems for the post-editing task.

The paralinguist cannot use the MT developers' automatic metrics, however, as decision factors to determine whether a candidate translation is of sufficient quality to be post-edited. This is because these MT developers' automatic metrics rely on comparison with reference translations. A paralinguist working operationally will not have the benefit of a reference translation to use these metrics. Therefore, part of this study will be a search for easily calculated metrics that do not require a reference translation yet yield indicators about a document's suitability for human post-editing.

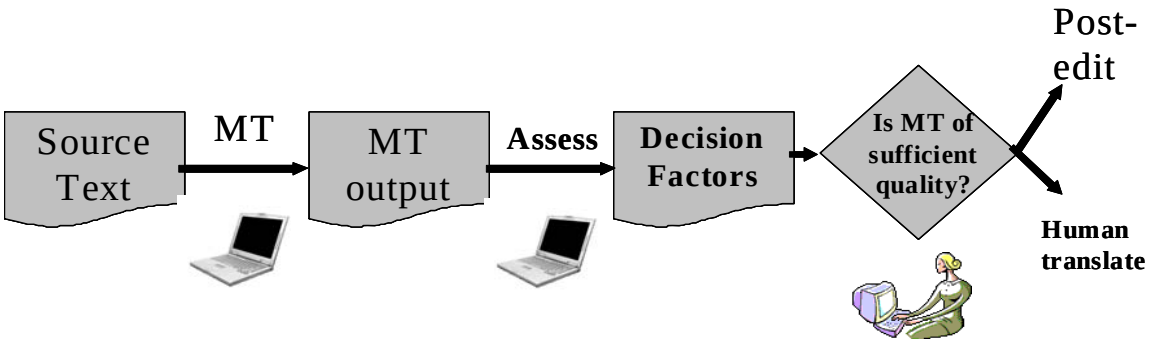

Figure 3. Paralinguist Assessment Data Flow

4 Investigation 2

The second investigation (Figure 3) answers the questions:

- Which decision factors aid a paralinguist in determining whether MT output is human post-editable? We hypothesize that such decision factors exist and that these can be automated.
- Are these decision factors measureable? Are they repeatable?
- Are they able to be generalized?
- What is the decision factors threshold above which the MT output can be used as a "seed translation?"

The process for answering these questions is:

1. We start with the MT output from the first investigation and use the error annotators scores determined during that investigation.
2. Candidate decision factors are analyzed for correlation with translator words per hour, translators' opinion of the post-editing activity and the error annotators' scores. It is important to note here that the decision factors apply to the document level as opposed to the sentence or segment level.
3. Those candidate decision factors that correlate well with translator words per hour, translators' opinion of the post-editing activity or the error annotators' scores are then selected for use in an operationally motivated test.

4.1 Potential Decision Factors

Given that MT developers' metrics cannot be used by a paralinguist in an operational setting, we search for easily calculated metrics that do not require a reference translation. Some early candidates include:

- Percentage of not-translated words;
- Language identification. Looking at whether the text identifies as the target language or the source language;

- Language model comparison. A language model comparison tests the fluency of a given translation by comparing that translation against a model of the output language. If the translation fits the language model, it is said to be fluent. Examples of this kind of metric include Corston-Oliver (et al., 2001), Blatz (et al., 2004) and Gamon (et al., 2005);
- Latent semantic analysis. Latent semantic analysis (LSA) is an information retrieval technique that enables two documents to be compared for similarity within a semantic space. In using LSA, the semantic space will be trained on a parallel corpus. Source and translation documents are then compared in this parallel space for similarity. Similar documents are considered to be adequate translations (Deerwester et al, 1990).; and
- Language Weaver TrustRank (Soricut & Echihabi, 2010). This is discussed below.

4.2 Language Weaver TrustRank

The Language Weaver TrustRank (also referred to commercially as TrustScore) is of special note here. Unlike the metrics described in the previous section, the TrustRank is used specifically to support decision making in a translation workflow. Instead of ranking translations from various systems, it calculates a category score for a given translation. this score can be used to determine if a document needs post-editing, retranslating or can be published as is.

As with other MT metrics (such as Gamon et al., 2005; Blatz et al., 2003; Corston-Oliver et al., 2001; Callison-Burch & Flournoy, 2001) TrustRank uses machine learning in combination with feature values to determine a final score. It calculates values for a number of features such as text-based features, language-model-based features, pseudo-reference-based features and example-based features. A model is then trained to predict the BLEU score-based ranking that would be given to the document if a reference translation were available, based on the values calculated for the features. TrustScore takes this one step further and is trained to predict user rankings based on the features.

These features of the TrustRank makes it attractive for the purposes of our study[3].

5 Case Study Specifics

We describe the specifics of the case study to include the data used, the participants and the systems we are using.

5.1 Data

The first study will examine Arabic-English, Chinese-English and Indonesian-English MT output. We chose these to show the range of maturity of MT systems. We selected documents from operational data where possible. The documents are at least 1000 words in length. They are as similar in difficulty as possible, but cover a wide range of domains.

5.2 Translators

The translation professionals and error annotators are members of the NVTC linguist resources cadre. The post-editors are source language native speakers. The error annotators are target language native speakers.

5.3 Systems

We refer to the MT system as MT System 1 for the purposes of this discussion. Additionally, the QuestSys question tool and the Morae key logging tool will assist in capturing the metrics on the translation process.

5.4 Location

We will perform the case studies in the NVTC Translation Technology Assessment Laboratory.

6 Conclusion and Future Work

We expect to present preliminary results at the conference. At the time of the writing of this paper, we are awaiting approval from the Institutional Review Board on our proposed use of human sub-

[3] We have asked Language Weaver to adapt TrustRank to accommodate a "post-edit or translate" decision factor for NVTC consideration.

jects in our study. We look forward to community input.

After the completion of the case study, we will work with additional MT systems and expand the pool of potential decision factors to include those traditionally used for selecting output in multi-engine MT systems, e.g., Nomoto (2003) and Callison-Burch & Flournoy (2001).

Acknowledgements

We wish to thank Daniel Marcu and the professionals at SDL Language Weaver for their insightful comments.

References

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

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, Cy. Kulesza, A., Sanchis, A., & N. Ueffing. 2004. Confidence Estimation for Machine Translation. COLING.

Callison-Burch, C. & R. Flournoy. 2001. A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines. MT Summit VIII.

Corston-Oliver, S., Gamon, M. & C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. ACL

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & R. Harshman (1990). "Indexing by Latent Semantic Analysis" Journal of the American Society for Information Science 41 (6): 391–407.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Proceedings of *Human Language Technology (HLT) Conference*.

Gamon, M., Aue, A. & M. Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. EAMT.

Lita, L., Rogati, M. & A. Lavie. 2005. BLANC: Learning Evaluation Metrics for MT. Proceedings of Human Language Technology Workshop on Empirical Methods in Natural Language Processing (HLT-EMNLP).

Nomoto, T. 2003. Predictive Models of Performance in Multi-Engine Machine Translation. MT Summit IX.

Papineni, K., Roukos, S., Ward, T. & W. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. *IBM Technical Report.*

Reeder, F. 2006. Direct application of a language learner test to MT evaluation. In Proceedings of AMTA 2006.

SAE. 2005. Translation Quality Metric. SAE International Technical Report, J2450.

Schütz, J. 1999. Deploying the SAE J2450 Translation Quality Metric in MT Projects. MT Summit VII.

Soricut, R., & A. Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In Proceedings of Association for Computational Linguistics (ACL).