Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input

Christine Bruckner
University of Munich, Germany
christin@cis.uni-muenchen.de

Mirko Plitt
Autodesk Development Sàrl, Neuchâtel, Switzerland
Mirko.Plitt@autodesk.com

Abstract

Following the guidelines for MT evaluation proposed in the ISLE taxonomy, this paper presents considerations and procedures for evaluating the integration of machine-translated segments into a larger translation workflow with Translation Memory (TM) systems. The scenario here focuses on the software localisation industry, which already uses TM systems and looks to further streamline the overall translation process by integrating Machine Translation (MT). The main agents involved in this evaluation scenario are localisation managers and translators; the primary aspects of evaluation are speed, quality, and user acceptance. Using the penalty feature of Translation Memory systems, the authors also outline a possible method for finding the "right place" for MT produced segments among TM matches with different degrees of fuzziness.

1 Introduction

The evaluation presented here was prepared during the MT Evaluation workshop held at the University of Geneva in April 2001. The ISLE taxonomy for the evaluation of Machine Translation (ISLE, 2001) was used as the basis for this evaluation task. The object of the evaluation was not to compare different MT systems or to find a formal measure for judging the quality of MT output, but rather to identify the implications of using MT in combination with other translation tasks.

This evaluation concentrates on the first part of the ISLE taxonomy (*1. Specifying User Needs*, especially *1.1 The Purpose of the Evaluation*, *1.2 The Object of Evaluation*, and *1.3 Characteristics of the Translation Task*) and in some aspects on section 2.2 (*System external characteristics*); section 2.1 (*System internal characteristics*), however, does not have much relevance for the scenario chosen here, in which general – and not system-specific – requirements are considered.

2 Evaluation Procedure

2.1 Purpose of the Evaluation

The purpose of this test evaluation was to develop a strategy for an *operational evaluation*[1] of the integration of MT output within the traditional software documentation translation process, which is typically based on the extensive usage of Translation Memory (TM) technology.

In a real-world scenario, the goal of such an evaluation would be to help a localisation manager decide whether or not MT output can be used as TM input in the process of localising software documentation.

2.2 Object of the Evaluation

"In the scenario of multilingual document production, MT has to be considered as part of a complex workflow [as it] has to interface with […] other processes" (Nübel & Schütz, 2000).

Machine Translation is evaluated here as part of a process involving both other translation technologies (here Translation Memories) and human intervention. Based on the ISLE taxonomy, the object of this evaluation is therefore an MT system considered as a component of a larger system (ISLE 2001, section 1.2.3).

2.3 Characteristics of the Translation Task

The output of the whole MT/TM translation process is to be used for *dissemination* purposes (ISLE 2001, section 1.3.2), which means that the translated (and post-edited) documents will be printed and/or published on the Web and thus made available to the end-users, i.e. used for *external publication* (ISLE 2001, section 1.3.2.2).

2.4 Specific Context of Use

The examined process represents an adaptation of the process typically followed by the software industry and the software localisation industry.

Software documentation tends to be highly repetitive and at the same time affected by frequent version updates. Therefore, TM technology has been used throughout this industry for a number of years. As a result, both software publishers and localisation agencies have built up large

[1] "According to White 2000, operational evaluations generally address the question of whether an MT system will actually serve its purpose in the context of its operational use." (ISLE, 2001, section 1.1.5)

corpora of translation memories, which are considered as important company assets.

Software documentation can be classified by *type* (e.g., tutorials, user manuals, programming references for developers), by *domain* (for example, office tools, CAD systems, business software), by *file format* (RTF, Framemaker, XML, HTML, and derived formats such JSP/ASP, etc.) and even by (software) *product*. See also section 1.5 *Input characteristics (author and text)* of the ISLE taxonomy (ISLE 2001), where the *Document type* (1.5.1), however, is only subdivided into *genre* (1.5.1.1) and *domain/field of application* (1.5.1.2).

In the particular context of software documentation, different products always require specific terminology, or at least glossaries with the corresponding software strings – often both, when Human Translation is combined with Machine Translation, like in the process examined here.

2.5 Who is the Evaluation Being Done for?

This is an additional consideration that is not explicitly mentioned in the ISLE proposal of April 2001, but is relevant for this evaluation task.

In this scenario, the outcome will be used by the localisation manager of a software publisher or by a project manager with sufficient decision rights.

The evaluation results would be equally interesting for his/her counterpart in a localisation agency, but the implications on the overall translation process may differ in that case.

2.6 Agents and Specific Needs

The following considerations are derivations and extensions of the aspects mentioned in section 1.4 of the ISLE taxonomy (*User characteristics*).

The principal aspect in which the localisation manager would be interested is whether or not the MT-enhanced translation process can shorten the "localisation delta" without any loss of quality. In addition to the mere speed criterion, the potential of automating the entire translation process by applying this kind of workflow could be determined. Another important aspect for the localisation manager are the cost benefits.

From the translators' point of view, the most important criterion is whether or not the usage of MT-produced segments will have a negative impact on the translators' efforts for doing their (TM-assisted) translation work. Factors to be considered here are the translators' status (in-house vs. free-lance, with direct or indirect/no access to additional resources, etc.), their TM expertise and MT experience (or openness to use MT). A benefit that translators should realize in any case from the MT-enhanced process is the increase in terminological consistency of the translations.

Reviewers and terminologists are less directly affected by the use of MT output in the TM translation workflow, but the possible impact on their specific task should nevertheless be considered in a real-life evaluation.

The TM expert in charge of fine-tuning the process will be particularly interested in identifying the most appropriate "penalty" for the MT segments that should be applied in the settings of the Translation Memory system. This penalty for MT segments – i. e., a percentage value indicating their divergence from a 100% match in the TM context – is a means to help the translator decide for a given translation unit whether a fuzzy TM match or the raw Machine Translation output should be preferred as a translation candidate (see also *2.9 A Method for Finding a Suitable TM Penalty Value for MT Segments*).

2.7 Evaluation Aspects Not Considered Here

The following aspects, although relevant in this context, are not considered here but should be addressed in a real-world evaluation. Except for the first two aspects, which represent system-internal characteristics, they could be sub-classified into section 2.2 of the ISLE taxonomy (*System external characteristics*).

- Comparison of different MT systems for usage along with a given TM system
- Language pairs supported both by the MT system and the TM system
- Costs of the MT system and the TM system (purchase prices, introduction and maintenance costs, training costs, cost savings, etc.)
- Quality and user-friendliness of MT/TM interface
- Possibility to export MT output directly into TM without alignment
- Potential to fully automate the import/export processes from and to the MT and TM system
- Impact of the MT input on the performance and stability of the TM system
- Additional TM maintenance needs
- Comparison of translation process with completely empty TMs (before MT import) vs. TMs previously filled with human translations (with "perfect" and "fuzzy" TM matches)

Most of these points should be determined by the individual localisation company (usually, these companies already use specific TM and/or MT systems and can decide best in which MT/TM systems the existing resources like terminology databases can be integrated, etc.). Such considerations have to be addressed during a pre-evaluation phase before the main evaluation.

2.8 Features to Be Evaluated

In this evaluation task, both qualitative and quantitative measures are used. For more detailed results, different weightings could be assigned to the individual measures in order to calculate an overall score.

The following features have been examined more closely and are ranked here according to their importance:

2.8.1 Speed

Description: Measure the time difference between a TM-based translation carried out *with* MT input and a TM-based translation performed *without* MT input. Does the import of MT segments really speed up the overall translation process?

Measure: Man-hours

Evaluation procedure: Define a test suite and have it translated by at least two teams of equally qualified translators: one team uses TM without MT input; the other team uses TM with MT input. Compare the time needed by each team to deliver a final translation of the test suite.

The test suite should consist of a collection of documents belonging to representative genres, domains, types and file formats. It could also contain different amounts of existing fuzzy TM segments and perfect matches in order to determine the respective gains in translation speed more exactly (e. g. are MT segments helpful in documents with many high-grade fuzzy TM matches?).

Score: Time required for carrying out the translation

Metric: Faster/Slower

See also section 2.2.4.1 of the ISLE taxonomy *(Time behavior)*, which is subdivided into *2.2.4.2.2 Production time/speed of translation*, *2.2.4.1.2 Reading time*, and *2.2.4.1.3 Revision and post-editing time (correction time)*.

2.8.2 Quality

Description: Measure the linguistic impact of feeding MT output in the TM-based process. Does it introduce more (terminological/stylistic) consistency, or do the translators tend to take over wrong/stylistically improper formulations from MT sentences?

Measure: Previously agreed quality standards such as the LISA QA Model defined by the Localisation Industry Standards Association (LISA, 2001)

Evaluation procedure: Apply the QA model to the test suites produced by the test teams, and compare the quality rating of each team on the basis of the QA system

Score: As defined in the QA model

Metric: *Better* or *Equal* (MT input does not deteriorate the translation quality)/*Worse* (MT input does deteriorate the translation quality)

See also section 2.2.1.2 of the ISLE taxonomy (*Accuracy*), which defines additional measures.

2.8.3 User Acceptance

Description: Measure in how far users (mostly translators) would accept or reject to work in a process that integrates MT output, and collect ideas how usability could be improved within the new process. If translators simply ignore MT produced segments, the whole MT integration process would be in vain.

Measure: User satisfaction

Evaluation procedure: Submit a questionnaire to the translators (asking them, for example, the following questions: Does the use of MT input make the translation work easier? Do you think that the quality and consistency of your translation improved? Is the use of MT a progress or a setback for the translation process?). Possibly assign different weightings to these questions according to their importance

Score: Results of the questionnaire

Metric: Acceptable/Unacceptable

2.9 A Method for Finding a Suitable TM Penalty Value for MT Segments

As mentioned above, penalties for MT segments in the Translation Memory settings are an important means to fine-tune the use of MT output when it is integrated in a Translation Memory system (see *2.6. Agents and Specific Needs*).

This is one of several (system and document) specific tasks in this complex workflow of MT and TM, for which an evaluation procedure should be exemplified in the following (for other tasks, see *2.7. Evaluation Aspects Not Considered Here*).

As each TM system uses its own algorithm for calculating the match percentages of TM fuzzy matches, different translation memory system show great divergences in their rating of fuzzy matches (for example, one TM system would assign a rate of 94% to a candidate segment, while another could would only assign a rate of 79%). Although a 70% threshold is often applied for retrieving candidates from the translation memory (i. e. only matches with 70% and above are shown to the user), it is not possible to make any statements about general thresholds values that would divide fuzzy matches into useful and useless translation candidates (see also Seewald-Heeg & Nübel, 1999).

Things are getting even more complicated when one tries to introduce a 'penalty' for MT matches, i. e. express their divergence from the "perfect" translation as a percentage rate. This penalty value also depends on the specific TM and MT systems in use and other factors involved in the translation process (languages, text type, etc.).

So the question here is at which percentage does a TM fuzzy match (i. e., a translated sentence – produced or revised by a human translator – that is stored in the translation memory and whose source sentence the TM system has found to be similar to the new source segment) prove to be more useful than a (non post-edited) MT-produced translation? The answer is essential for ranking TM and MT translation candidates in the user interface of the TM system and, in consequence, for setting the minimum percent value at which fuzzy matches should be retrieved and presented by the TM system. (For example, it would be nonsense to apply a 35% penalty to MT segments when only matches above 70% are retrieved.)

The question to be answered is whether a fuzzy match with a score of e. g. 80% (calculated by the TM system) is more useful than a translation generated by the MT system (100% – 15% MT penalty = score of 85%; where the MT penalty is a user-definable value, and is precisely the value that is be determined in this specific evaluation task).

In order to find the optimal penalty value *x*, a variation of the "edit distance"[2] could be used as measure. The edit distance is often used in evaluating MT post-editing and is obtained by counting the steps required to bring the translation initially suggested up to the desired quality.

[2] "Edit distance counts the total number of 'insert, delete and swap order' operations (all other are broken down into these three)." (ISLE 2001, comment to section 2.1.4.1 *Post-editing or Post-Translation)*

So, in this evaluation scenario, the edit distance can be measured by the number of editing operations the translator has to perform in order to produce an acceptable and correct translation out of the translation suggested by the TM (which may be a fuzzy match retrieved from the human translations in the TM or an imported MT segment to which a penalty value has been applied).

By comparing the work involved in editing different fuzzy TM matches with different percentage values and the work that is required to bring MT produced translations to the desired translation quality, it is possible to define a close-to-optimal threshold up to which MT translations should be preferred to fuzzy TM matches. Such threshold would also help to achieve the first goal of this evaluation scenario, which is *speed* (see section 2.8.1), as the user would always be given the translation candidate (TM match or MT translation) that requires the least amount of editing work.

2.9.1 Experiment

Languages: English (source), German (target)

Document type: Software documentation

Systems used: Customised Systran system (MT), Trados Translator's Workbench ™

Description: The Trados TM contained perfect and fuzzy matches from previous translations. 100% matches had been pre-translated in the new documents (with the Trados "Pre-translate" function), and only the segments with a match level of 99% and below were exported to Systran. The MT output from Systran was aligned and imported into the translation memory. (The alignment step is not necessary in the standard Systran system as it offers direct export facilities for Trados. But here a customised Systran system with a Web interface was used, which already contained customer-specific adjustments concerning terminology, style, etc. to enhance the quality of the Systran translation for the company's documents.)

By default, the TRADOS Translator's Workbench applies a penalty of 15% to MT segments. The goal of this experiment was to find out whether this penalty was sufficient for the MT-produced segments (i. e., is a 85% rated MT match really "better" than a 84% or lower-grade fuzzy TM match?). Although the TRADOS system can display all fuzzy matches above a user-defined threshold, only the highest match is actually presented to the translator; the matches with lower percentage values are hidden and only available by clicking on buttons in the TM user-interface. Moreover, the highest ranked match is usually automatically copied into the translator's editor, so additional editing efforts are required when the translator wants to substitute a lower ranked, but in his/her opinion more suitable, match for the pre-inserted translation suggestion. This makes it important to assign an adequate penalty to MT segments, otherwise the fuzzy matches with lower match values but higher translation quality may be suppressed.

The object of this experiment was not to investigate a large number of sentences in order to find the perfect threshold rate for the test suite, but rather to demonstrate this approach and see whether it was feasible at all in a real world scenario. For this reason, 4 suitable source sentences were selected, which already contained fuzzy matches of different percentage rates:

For example 1, there were fuzzy matches of 95%, 89%, 53%, 48%, and the MT match (rated at 85% according to the 15% Trados penalty).

For example 2, there was a fuzzy match of 76% and the 85% MT match.

For example 3, there was a fuzzy match of 71% and the 85% MT match.

For example 4, there were fuzzy matches of 76% and 66% and the 85% MT match.

Evaluation procedure: Count word deletions, word insertions and word changes (both changes in the word position and changes in morphology) that are necessary to bring the MT output to the desired quality, and compare the result with the number of corresponding operations required for editing different fuzzy matches (in our example the 89% match, the 76% match, the 71% match, the 64% match, etc.).

2.9.2 Results of the Experiment

From the examined examples it could be concluded that the standard penalty of 15% that TRADOS Workbench assigns to MT-based segments is too low for the test suite: For example, the 76% fuzzy match rated required less post-editing than corresponding MT match.

For a thorough and reliable evaluation study, these few examples are, of course, not enough; the number of test sentences should be much higher. More source sentences of different length and complexity (and probably format changes) and their corresponding TM fuzzy matches and MT translations should be examined. However, it will often be a problem to find comparable test sentences with existing fuzzy matches in the adequate range between 60 and 90%. As we are dealing with real-world evaluations in the software localisation industry, we would not advise to construct artificial corpora of similar sentences (as this is usually done in TM and MT evaluations), but rather rely on fuzzy matches from real-world corpora.

As an alternative, the time needed for editing TM fuzzy matches and MT produced matches could be measured in order to find a suitable threshold. Yet this would involve many user-specific aspects, and it would be difficult to come to an objective result.

The experiment has also shown that the three categories – insertions, deletions, and changes – are not fine-grained enough: It would be useful to differentiate between a morphological change and a change in the sentence position. Moreover, the insertion of several consecutive words that all belong to one noun phrase or prepositional phrase should be weighted differently than the insertion of the same number of separate words that are spread all over the sentence; the same applies to similar problems.

2.9.3 Related work

Carl and Hansen (1999) have shown in a similar, yet more theoretical study that above a (previously fixed) threshold of 80%, the quality of the matches given by translation memory systems (Trados, Transit, and Zeres) is better than the quality of the translations produced by their

example-based machine translation system (EDGAR). In order to compare fuzzy matches and MT output, they calculate translation scores based on the common number of words/lexemes in the ideal translation and in the fuzzy match or MT translation, respectively. Although this is a more mathematically sophisticated approach to the problem of finding a suitable threshold, it does not give a satisfactory measure for the actual amount of real work and time that the user (translator) needs to transform such translation candidates into "perfect" translations.

3 Conclusions

This paper presents a procedure for evaluating the integration of machine-translated segments into the Translation Memory systems, using the software localisation industry as a real-world scenario. We come to the conclusion that for such scenario, the cost factor is usually not the most important aspect, but the focus is rather on higher translation *speed*, on the preservation/improvement of translation *quality* and on the *acceptance* of this MT/TM combination by the end-users (usually translators). For an effective integration of MT segments into TM systems, it is, among other things, crucial to determine an adequate "penalty" value for MT segments in order to rank them adequately against fuzzy matches and provide the user with the translation candidate that requires least post-editing efforts.

The presented considerations and experiments are based on a practical scenario with existing MT and TM systems, document types, languages involved, etc. It should be pointed out once again that there is no ideal MT/TM combination for all purposes, but such decision always depends on the specific conditions and requirements of the company's environment.

4 References

Carl, M. & Hansen, S. (1999): Linking Translation Memories with Example-Based Machine Translation. http://www.iai.uni-sb.de/~calr/iaiwp/p7/index.html

EAGLES (1999). EAGLES 7-step Recipe for NLP Evaluation. http://issco-www.unige.ch/projects/eagles/ewg99/7steps.html

ISLE (2001). ISLE Taxonomy for MT Evaluation. http://issco-www.unige.ch/projects/isle/taxonomy2/

LISA (2001). The LISA QA Model. http://www.lisa.unige.ch/products/qamodel.html

Nübel, R. & Schütz, R. (2000). Evaluation as Language Technology Deployment Trigger. Paper presented at the EAMT Workshop, Ljubljana, May 2000.

Seewald-Heeg, U. & Nübel R. (1999): Ausblick: Evaluierung von Translation-Memory-Systemen. In: LDV-Forum Bd. 16 (1999) 1/2.