Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input in the Translation Process of Software Documentation

(Preliminary Version)

Christine Bruckner, University of Munich, Germany (christine.bruckner@gmx.de)
Mirko Plitt, Autodesk Development Sàrl, Neuchâtel, Switzerland (Mirko.Plitt@autodesk.com)

▪ Purpose of the Evaluation

The purpose of this evaluation was to develop a plan for the *operational evaluation*[1] of the integration of MT output within the traditional software documentation translation process 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 documentation localisation process.

▪ Object of the Evaluation

Machine-Translated segments within a TM-based translation workflow can be considered as *An MT system considered as a component of a larger system*[2].

▪ Specific Context of Use

The process evaluated is specifically from a standpoint of the software industry and the software localisation industry.

Software documentation tends to be highly repetitive yet 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 localization agencies have built up large corpora of translation memories, considered as important company assets.

Software documentation can be classified by type (e.g., tutorials, user manuals, programming references for developers), domain (e.g., word processing, CAD, financial software) and even product, as specific products always require specific terminology, be it within the context of Machine Translation or Human Translation, or both, as in the process examined here.

▪ Who is the evaluation being done for?

Localisation management of a software company

▪ What agents and specific needs have to be considered?

Software company managers: speed, quality, automation potential of entire translation process

Localisation project managers: speed, quality

The main decision criterion in this context is the gain in production time without quality loss. Other important aspects for the localisation manager are the cost-saving potential and the increase of flexibility through possible process automation.

Translators (internal/external; TM experience, no MT experience – same needs for all translators): make translation work easier, improve consistency

[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. The primary factors include the cost-benefit of bringing the system into the overall process (costs)." – ISLE framework draft 1.1.5
[2] ISLE framework draft 1.2.3

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 effort required to carry out their work.

Reviewers

(Terminologists: integration of new terminology both in term banks and in MT system)

- **Possible tasks:**

Find out which MT system gives the best quality output for TM input: -> comparison of several MT systems; not considered in this evaluation

Given a specific MT system: compare the translation process using a translation memory filled with MT translated segments with the translation process using an empty translation memory or TM containing fuzzy matches

Given a specific MT system and a given TM system with a translation memory containing 100% matches and fuzzy matches: which penalties should be reasonably applied to MT produced segments?

Other considerations (not treated here):

Are the required languages covered by both the MT and the TM system?

Does the selected TM system offer an interface to a specific MT system (interactive translation – batch translation)?

Does the TM system offer a pre-analysis function for 100% matches, fuzzy matches, repetitions?

Is it possible in the TM system to do a pretranslation using the TM, then export segments below a certain match threshold and have only such segments translated by the MT system?

Is alignment of the MT output required or does the MT system offer a direct export into a given translation memory system/the TMX format?

How much work is needed to automate the MT export/TM import process?

- **Features to be evaluated:**

FEATURE	DETAILS	MEASURE	EVALUATION PROCEDURE	SCORE
1. Speed (time-to-market, sim-ship in software localisation)	Measure the difference between TM with MT input and TM without MT input	Man hours	set up 2 teams of translations: 1 team uses TM without MT input, 1 team uses TM with MT input =>compare the time needed by each team Test suite with representative documents: - genre: technical documentation - domain: software - different document types (manuals, online help, web pages, etc.) - different file formats (rtf, html, mif, etc.) apply different amounts of existing fuzzy TM segments (e.g. 80% existing TM matches and 20% MT segments vs. 20% existing TM matches and 80% MT segments) – matches may be only perfect matches or perfect and fuzzy matches	Is the translation with MT input quicker than the translation with TM only? - **Yes/No**
2. Quality	Does the translation quality deteriorate when MT translations are suggested? Does the (terminological) consistency of the translated documents improve when suggestions from the customized MT system are available?	Use of given QA system for human translations used in the localisation industry (LISA): error rate, style, etc.	Set up 2 teams of translations: 1 team uses TM without MT input, 1 team uses TM with MT input =>compare the errors of each team based on the QA system	**better – equal – worse**
3. End-user acceptance	Do the translators use the MT segments at all? Do they think the review of MT segments is more work than a translation from scratch? Do they think the MT segments are helpful? Are they satisfied with this kind of translation	User satisfaction	Questionnaire: - Does the use of MT input make the translation work easier? - Do you think that the quality and consistency of your translation is improved? - Do you think the use of MT is a progress or a setback for the translation process?	Easier – cannot say – more difficult Improved – cannot say – worse Progress – cannot say – setback -> possibly apply a weighting

FEATURE	DETAILS	MEASURE	EVALUATION PROCEDURE	SCORE
	process?			to the different questions
4. Costs	Introduction cost of MT system, additional maintenance cost of TM system, developing/introduction cost for new utilities (input/output converters etc.) Cost savings by lower word prices for MT-pretranslated segments (-> effects on external translators); higher throughput: less translators needed (effects on internal translators)	Depending on MT system/TM system used, company, use of internal/external translators; not evaluated here	Not evaluated here	Not evaluated here
5. Influence on performance/ stability of TM system	TM sizes are growing/exploding by addition of MT segments: possible loss in performance or stability of TM system	Depending on TM system and size of the memories; not evaluated here	Not evaluated here	Not evaluated here
6. Maintenance	Additional work needed for maintaining/cleaning of TMs	Depending on TM system, amount of “superfluous” MT segments; not evaluated here	Not evaluated here	Not evaluated here

=> assign weightings to the individual scores in order to find an overall score?

- **Which "penalty" (in percent) should be applied to MT segments?**

The match percentage of fuzzy matches always depends on the *specific TM system* used (every TM system uses its own fuzzy matching algorithm):

At which percentage is a TM fuzzy match (produced/revised by a human translator) more useful than an MT produced translation (important for the ranking of alternatives/setting the minimum match value)?

Example: fuzzy match with 80% score (calculated by the TM system) vs. MT segment (100% - X% penalty for MT = Y% score; where X is user definable)

"Edit distance" is sometimes to measure the number of steps in order to bring the MT output up to an acceptable quality[3]. Actually, the fuzzy matching algorithm of TM systems is often based on the calculation of edit distance.

It could be useful to apply this measure to TM fuzzy matches of different percentages values (30%, 40%,…, 99%) and to their MT counterparts in order to find out at which threshold TM fuzzy matches should be ranked higher than MT produced segments.

Experiment:

Languages used: English -> German

Document type: software documentation

Systems used: (customized) S1 system, S2 Translator's Workbench

The S2 TM already contained "perfect" and "fuzzy" matches from previous translations. 100% matches had been pre-translated in the new documents, and only the segments with a match level below 99% were exported to S1. MT output from S1 was aligned and imported into the translation memory (this alignment step is not necessary in the standard S1 system as it offers direct export facilities for S2).

By default, the S2 Translator's Workbench applies a penalty of 15% to MT segments. The goal of this experiment was to find out whether this penalty was enough for the MT produced segments. Although the S2 system gives all fuzzy matches above a user-definable threshold, only one match is actually shown, and the translator has to click through the TM window in order to find matches with lower values (usually, the highest ranked match is copied into the editing window, so inserting lower ranked matches requires additional editing time).

Evaluation procedure: Count the number of word deletions, insertions and (position/morphological) changes that are needed to bring the MT output to a good quality and compare it to the number of word *deletions*, *insertions* and (position/morphological) *changes* needed to bring fuzzy matches (with different match values) to a good quality.

Example 1:

Source:
The dialog box closes temporarily, and XXX YY prompts you to select objects.

Human translation for reference purposes *(not provided by MT nor by TM)*
Das Dialogfeld wird vorübergehend geschlossen, und XXX YY fordert Sie auf, Objekte auszwählen.

95% fuzzy match:
The dialog box closes temporarily, and XXX prompts you to select objects.
Das Dialogfeld wird vorübergehend geschlossen, und XXX YY fordert Sie auf, Objekte auszwählen.

- insert 1 word

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

89% fuzzy match:
The dialog box closes temporarily, and XXX YY prompts you for object selection.
Das Dialogfeld wird vorübergehend geschlossen, und XXX YY fordert Sie zur Auswahl von Objekten auf.

- no changes required (only different grammatical structure in source and target sentence)

85% MT match (15% default penalty):
Das Dialogfeld wird ~~schließt~~ vorübergehend geschlossen, und XXX YY~~Anfrage~~ fordert Sie auf, ~~um der~~ Objekte auszuwählen.

- deletion of 4 words
- insertion of 4 words
- change of 1 word

53% fuzzy match:
Closes the dialog box temporarily so that you can select objects in your drawing.
~~Schließt~~ ~~d~~Das Dialogfeld wird vorübergehend geschlossen, ~~damit~~ und XXX YY fordert Sie auf, Objekte ~~in der Zeichnung~~ auszuwählen ~~können~~.

- deletion of 6 words
- insertion of 7 words
- change of 3 words

48% fuzzy match:
Closes this dialog box and opens the selected file in XXX YY.
~~Schließt~~ ~~d~~Das Dialogfeld wird vorübergehend geschlossen, und XXX YY fordert Sie auf, ~~öffnet die gewählte Datei in XXX YY~~ Objekte auszuwählen.

- deletion of 6 words
- insertion of 8 words
- change of 3 words

Example 2:

Source:
The name can have up to 255 characters and can include letters, numbers, blank spaces, and any special character not used by <XXXXXXXXXXXXX> and XXX YY for other purposes, if the system variable EXTNAMES is set to 1.

Human translation for reference purposes *(not provided by MT nor by TM):*
Wenn die Systemvariable EXTNAMES auf 1 gesetzt ist, kann der Name bis zu 255 Zeichen umfassen (Buchstaben, Ziffern, Leerzeichen sowie Sonderzeichen, die nicht bereits in <XXXXXXXXXXXXX> oder XXX YY für andere Zwecke belegt sind).

85% MT match (15% default penalty):
Der Name kann bis zu 255 Buchstaben ~~haben~~ enthalten und kann aus Buchstaben, ~~Anzahl~~ Ziffern, ~~unbelegte Raum~~ Leerzeichen und ~~jeden speziellen Buchstaben~~ Sonderzeichen ~~mit einschließen~~ bestehen, ~~der~~ die nicht ~~durch~~ von <XXXXXXXXXXXXX> und XXX YY für andere Zwecke verwendet ~~wird~~ werden, wenn die ~~SYSTEMVARIABLE~~ Systemvariable EXTNAMES auf 1 eingestellt ~~wird~~ ist.

- deletion of 13 words
- insertion of 11 words

- change of 2 words/positions (SYSTEMVARIABLE, position of conditional clause)

76% MT match:
View names can have up to 255 characters and can include letters, numbers, blank spaces, and any special character not used by <XXXXXXXXXXXXX> and XXX YY for other purposes.

Wenn die Systemvariable EXTNAMES auf 1 gesetzt ist, kann ~~Dd~~er Name ~~kann~~ bis zu 255 Zeichen umfassen (Buchstaben, Ziffern, Leerzeichen sowie Sonderzeichen, die nicht bereits in <XXXXXXXXXXXXX> oder XXX YY für andere Zwecke belegt sind).

- deletion of 0 words
- insertion of 13 words (8 for conditional clause, 3 for prepositional phrase, 2 trademark symbols)
- change of 2 words/positions (Der -> der; position of “kann”)

Example 3:

Source:
XXX YY always checks for Object Enablers regardless of your settings in the Today window.

Human translation for reference purposes *(not provided by MT nor by TM):*
XXX YY sucht immer nach Objekt-Aktivierern, und zwar unabhängig von den Einstellungen im Fenster Aktuell.

85% MT match (15% default penalty):
XXX YY ~~Aktivieren~~ sucht immer ~~für~~ nach Objekt~~-Enablers-~~Aktivierern, unabhängig ~~davon-~~von Ihren Einstellungen im ~~Tag~~Fenster Aktuell.

- deletion of 5 words
- insertion of 5 words
- change of 1 word

71% match:
Prevents XXX from checking for Object Enablers regardless of your settings in the Today window (see TODAY).
~~Verhindert, daß~~ XXX YY sucht immer nach Objekt-Aktivierern ~~sucht~~, und zwar unabhängig von Ihren Einstellungen im Fenster Aktuell ~~(siehe auch AKTUELL).~~

- deletion of 5 words
- insertion of 2 words
- change of 1 words/positions

Example 4:

Source:
Sets the units XXX YY uses for an object being inserted into the current drawing when no insert units are specified with the INSUNITS system variable.

Human translation for reference purposes *(not provided by MT nor by TM):*
Legt die Einheiten fest, die XXX YY für ein Objekt verwendet, das in die aktuelle Zeichnung eingefügt wird, sofern keine Einfügungseinheiten mit der Systemvariablen INSUNITS festgelegt wurden.

85% MT match (15% default penalty):

~~Stellt~~ Legt die Einheiten ~~ein~~ fest, XXX YY~~, dasgebrauch~~ für ein Objekt verwendet, das ~~ist,~~ in ~~der~~ die aktuelle Zeichnung eingefügt~~en~~ wird, wenn keine ~~EinsatzEinheit~~ Einfügungseinheiten mit der Systemvariablen INSUNITS ~~SYSTEMVARIABLE~~ spezifiziert w~~e~~urden.

- deletion of 9 words
- insertion of 5 words
- change of 4 words

74% match:
Sets the units XXX uses in the current drawing when no insert units are specified with the INSUNITS system variable.

Legt die Einheiten fest, die XXX YY ~~in der aktuellen Zeichnung verwendet~~, für ein Objekt verwendet, das in die aktuelle Zeichnung eingefügt wird,sofern keine Einfügungseinheiten mit der Systemvariablen INSUNITS festgelegt wurden.

- deletion of 1 word
- insertion of 8 words
- change of 4 words

66% match:
Sets which units to automatically use in the current drawing when no insert units are specified with the INSUNITS system variable.

~~Richtet die automatisch in der aktuellen Zeichnung zu verwendenden Einheiten ein~~, Legt die Einheiten fest, die XXX YY für ein Objekt verwendet, das in die aktuelle Zeichnung eingefügt wird, wenn keine Einfügeeinheiten mit der Systemvariablen INSUNITS angegeben wurden.

- deletion of 7 words
- insertion of 14 words
- change of 5 words/positions

Result:

The standard penalty of 15% that S2 Workbench assigns to MT-based segments is too low for the tested fuzzy matches: For example, fuzzy matches with 76% require less post-editing than MT matches

For the test texts, a penalty of about 30% should be applied, i. e. MT-based segments should be presented as 70% matches, and therefore be ranked below fuzzy matches with higher percentage values.