Uncontrolled Terminology and MT: The Importance of Making Good Comparisons

by Rafael Guzmán

(Originally published in *Translation Journal*, vol. 13, No. 2, April 2009)

Key words: machine translation, uncontrolled terminology, duplicates, comparison criteria

Introduction

Feeding Machine Translation (MT) dictionaries with uncontrolled terminology is never a good thing. When this is done, MT engines consistently produce translations containing the same unwanted or wrong terminology over and over again. Post-editing these terminology errors tends to be time-consuming and expensive, which defeats the purpose of using MT.

Feeding Machine Translation (MT) dictionaries with uncontrolled terminology is never a good thing.

One of the most effective ways of controlling MT terminology is to automatically check for the existence of terms within an MT dictionary and across multiple terminology repositories, and determine if they are duplicated and how they are translated. This type of control can involve extensive terminology comparisons. Unfortunately, there is still a broad unawareness of why and how appropriate comparison criteria need to be used for each specific situation. This is confirmed by the lack of sufficient comparison functionality and flexibility available in current terminology tools.

This article briefly explains why terminology for MT dictionaries needs to be controlled, and suggests necessary comparison criteria as well as why Excel spreadsheets can help in filling in the gaps in current terminology tools to automate and customize terminology comparisons. Although the emphasis of this article is on terminology tasks affecting rule-based machine translation, they can be applicable to terminology management in general.

Good-quality terminology: nice to have or critical?

A while ago, Client Side News (2006) warned that "businesses often fail to see terminology management as a way to cut costs," and Warburton (2005) pointed out that "terminology is the biggest factor in poor translation quality." Also, it has been discussed the "unexpected return on investment (ROI)" generated by good-quality terminology (Wittner, 2007).

These statements seem to be confirmed by Lionbridge (2009), which reports that "it is estimated that 15 percent of all globalization project costs arise from rework and that the primary cause of rework is inconsistent terminology."

This poses the question of whether the success of rule-based MT is exempt from the dangers of uncontrolled terminology.

Unfortunately, it is not. Because rule-based MT relies on dictionaries, apart from grammar rules, the danger is twofold:

1. Necessary terminology is missing
2. Encoded (appropriate or inappropriate) terminology is incorrectly or inconsistently translated

The first problem is the result of poor terminology harvesting and validation, while the second problem is caused by the lack of appropriate quality controls. This article will focus only on the latter.

Once incorrect translated terminology makes its way into an MT dictionary, the same wrong translations will automatically be repeated over and over again in every new MT output.

This is caused by the fact that MT dictionaries are usually expected to contain just one single instance of each term and one unique translation and part of speech (POS)—maybe some contextual information as well—in each semantic domain. If there are terms that are duplicated and translated inconsistently, the MT engine will not know which translation to use. Consequently, it is possible that the wrong translation may always end up being used in every MT output. But even if there are no duplicates, or if they are translated consistently, the MT engine will use the wrong translation if terms contain an unwanted or wrong translation.

The severity of this problem will normally depend on the number of terms affected and their frequency in the source text. This suggests how critical it can be that necessary terminology be captured and translated correctly in the MT dictionary.

To sum up, post-editing incorrectly translated terms in the MT output is a reactive approach that causes unnecessary extra costs. What is needed is a proactive terminology control strategy that will prevent MT from causing the exact same problems over and over again.

MT-related terminology tasks

A very effective strategy to control MT terminology is to automatically analyze MT dictionaries and check for:

1. Duplicated source terms. If found, are they translated consistently?
2. Key source terms occurring as head of compound nouns. If found, are they translated consistently?
3. Deprecated (unwanted) or wrong translations

There are numerous MT-related terminology tasks in which these controls are required. These controls always involve terminology comparisons within an MT dictionary and across multiple terminology repositories (Table 1).

Table 1: terminology tasks involving comparisons

MT-related terminology task	Reason why a comparison task is required
Validation of harvested terminology for an MT dictionary	Once harvested terms have been manually "cleaned," they need to be checked for duplicates.
Creation of a global list of key terms from different project glossaries to feed an MT dictionary	Once all the project glossaries are merged into a single glossary, terms need to be checked for unwanted terms, duplicates, and translation inconsistencies.
Populating a list of source terms by leveraging translations and other data from legacy glossaries to create an MT dictionary	Before automatically allocating translations to each term, their part of speech values (POS) values need to be compared to avoid false positives (e.g. "open" can be a verb or an adjective; "scan" could be a verb or a noun) .
Customization of a project glossary into an MT dictionary from scratch	Some rule-based MT dictionaries often require that terms are surrounded by linguistic clues based on their POS values (e.g. open > **to** open). This requires comparing POS values as a pre-condition to allocate the appropriate clues.
Feeding an MT dictionary with terminology from a project glossary	Some of the terms in the glossary may already exist in the MT dictionary with the same or a different translation.
Diagnosing (auditing) an MT dictionary	Terms may contain unwanted translations. There can also be duplicates translated inconsistently.
MT dictionary metrics	Based on any of the tasks above, metrics need to capture the number of duplicates, deprecated translations and inconsistencies

Finally, when dealing with MT dictionaries, "it is not wise to hope that all source texts have been written well, and have followed any type of standardization guidelines with respect to

the use of terminology grammatical rules, or writing style" (Allen, 2006). Sometimes it may be necessary to encode incorrect source terms (e.g. deprecated or misspelled terms) that tend to recur frequently, otherwise, they will not be recognized by the MT engine. However, a better solution would be to use a normalization dictionary. What really matters is to ensure that their translation in the dictionary is correct and consistent.

Terminology comparison criteria

In the previous sections, the importance of checking for duplicates as well as for inconsistent and deprecated translations in MT dictionaries has been discussed. But is this enough? In other words, is there anything that can invalidate an automatic terminology comparison or make it inaccurate causing the issues to remain hidden?

The tables below show some basic self-explanatory examples of likely results after comparing terms automatically using specific criteria. Each table represents a different comparison scenario. The equal and unequal signs indicate the result of the comparison. Of course, many of these scenarios often appear combined in practice.

Table 2: Spelling

Scenario	**Example**
Same term, same spelling	administrator = administrator
Same term, but different spelling	administrator ≠ admin

Tables 3 and 4 represent two of the most typical comparison scenarios. Suppose that a rule-based MT dictionary (e.g. Systran) requires that generic terms are entered in lower case in order to enable the engine to recognize them in both lower and upper case while translating the source text. If a term is entered in upper case only, the MT engine will only be able to recognize the exact occurrence of that term in upper case.

Based on this, if the term "Virtual Memory" (upper case) already exists in the MT dictionary and the term "virtual memory" (lower case) is compared against it using a case-insensitive comparison, it will be "wrongly" concluded that "virtual memory" already exists in the dictionary and that there is no need to add it again. On the other hand, if a case-sensitive comparison is run, "virtual memory" will be flagged as an non-existent term in the dictionary. But if this term gets added to the dictionary in this case, will this action not create a duplicate?

Table 3: case sensitive versus case insensitive

Scenario	Example
Same term, same case	virtual memory = virtual memory
Same term, but different case	Virtual Memory ≠ virtual memory

Table 4: Part of speech (POS) values

Scenario	Example
Same term, same POS	scan (noun) = scan (noun)
Same term, different POS	scan (noun) ≠ scan (verb)
Same term, but POS missing in one of them	scan (noun) ≠ scan
Same term, same POS spelled differently	fast (adj.) ≠ fast (adjective)

Using POS values can help to fine-tune terminology comparisons. For instance, proper nouns are normally (although not necessarily) expected to be encoded in upper case, while nouns, verbs and adjectives are expected to be encoded in lower case. Occurrences of the same term with different part of speech values are not considered duplicates. For instance: "Virtual Memory" (proper noun) and "virtual memory" (noun) could be safely added to the MT dictionary.

When POS values are not used as part of the comparison criteria, misleading results are likely to occur.

Obviously, this comparison criterion assumes that all the candidate terms for addition to the MT dictionary, as well as the existing terms in the dictionary, have the correct spelling and correct standard POS values (e.g. TBX or OLIF). This might require a little bit of manual "cleaning" before hand.

Table 5 shows some scenarios of possible false positives if a "match whole word" option is not enabled prior to the comparison.

Table 5: Whole word matching

Scenario	Example
Term fully matches a term	backup = backup
Term partially matches the same term	backup = backups
Term partially matches a different term	backup = BackupExec

Some terms surrounded by "noise" (e.g. linguistic clues, trade mark sign) may need to be added to the MT dictionary even if the same term already exists without noise. In many cases, they do not need to be treated as duplicates necessarily; otherwise they will not be recognized by the MT engine (e.g. Linux, Linux(tm)). In these cases, what really matters is that the translation remains consistent and correct. To check this, any existing "noise" may need to be temporarily removed before running the comparison analysis.

Table 8: "Noise" (linguistic clues, punctuation, ampersand, tags, TM, variables)

Scenario	Example
Same verb, preceded inconsistently by "to"	open ≠ to open
Same term, but linguistic clues follow one of them	offline ≠ offline (flexible hyphen) off-host ≠ "off-host" (a)
Same term, but asterisk indicating the existence of a comment follows one of them	backup ≠ backup (*)
Same term, but colon sign follows one of them	Space available ≠ Space available:
Same term, but trade mark (TM) sign or similar appended to one of them	Linux ≠ Linux(tm)
Same term, but hotkey in the middle of one of them	File ≠ F&ile
Term surrounded by tags	Activate ≠ <tag>Activate</tag>

Table 6: semantic domain values

Scenario	Example
Same term, same meaning in different O/S	file (in Windows) = file (in Macintosh)
Same term, different meaning in different O/S	Dock (in Windows) ≠ Dock (in Macintosh)

Table 7: Leading and trailing blank spaces

Scenario	Example
Same term, no leading or trailing spaces	CPU Usage = CPU Usage
Same term, but leading space	CPU Usage ≠ CPU Usage

Table 9: translations

Scenario	Example
Same term, same translation	scan (análisis) = scan (análisis)
Same term, different translation	scan (análisis) ≠ scan (escaneo)

Excel spreadsheets as terminology tools companion

So far, this article has discussed why rule-based MT needs terminology controls focused on terminology comparisons using appropriate criteria based on the relevant MT terminology task. However, carrying out these tasks manually is tedious, time-consuming and prone to errors.

This raises one more question: what is the best technology environment to run good terminology comparisons?

While it is true that a number of well-known terminology tools have improved significantly providing online storage and management capabilities, the comparison functionality

described in this article is still very limited in many tools. In addition, there is also the lack of flexibility to customize and automate the terminology tasks required by each situation. Also, practice shows that terminology stored in powerful terminology repositories often ends up being exported to text files that are subsequently opened in spreadsheets for evaluation or even usability purposes.

This is the context in which Microsoft Excel spreadsheets can be of great assistance to terminology tools, not necessarily as storage repositories but as terminology workbenches for analysis and customization purposes.

In Excel, users can easily record macros and develop them further in the Excel Visual Basic (VB) editor according to their own needs (e.g. T-Manager Terminology Tool).

For instance, the VB editor allows users to instruct a macro to compare all the cells in a dictionary or glossary in a spreadsheet and flag each duplicated term using the cell next to each duplicate. Then the user can start adding more macros and features such as a drop-down list in one of the cells in the spreadsheet to instruct the macro to use the required comparison criterion (e.g. case-sensitive or case-insensitive).

Obviously, apart from a bit of creativity, some previous knowledge of VB will be also needed, but there are plenty of online tutorials and this can be easily learned as the user goes along.

Conclusion

Good-quality and controlled terminology is critical for the success of rule-based Machine Translation. Key to this is to automatically identify duplicates and translation inconsistencies, as well as deprecated translations within an MT dictionary and across multiple terminology tools. This requires making sure that the necessary terminology comparisons are made using appropriate comparison criteria. As a result, MT outputs will require less post-editing, and MT translation costs will be reduced. Finally, while many current terminology tools provide little or no comparison functionality and flexibility, Excel macros and spreadsheets can help to provide these.

Acknowledgments

The author thanks Dr Johann Roturier and Fabio Fantino for discussion and useful comments.

References

Allen, J., 2006. Improved translation quality with Machine Translation Dictionary Building. [Online]. Available at: http://www.translatorscafe.com/cafe/article59.htm [accessed 20 November 07].

ClientSide News. 2006, A New Mid-level Solution for Terminology Management. [Online]. Available at: http://www.lexicool.com/lingo4-terminology-management-article.asp?IL=1 [accessed 20 January 2008].

Guzmán, R., 2007. T-Manager Terminology Tool. [Online]. Available at http://www.invenis.net/resources/tw/index.php (Accessed 18 February 2009].

Guzmán, R. 2008. Advanced automatic MT post-editing. *Multilingual,* 9 (3), p.52-57.

Itagaki, M., Aikawa , T, and He, X., 2007. Automatic Validation of Terminology Translation Consistency with Statistical Method.[Online]. Available at: http://research.microsoft.com/en-us/um/people/xiaohe/publication/mtsummit-2007.pdf [accessed 2 February 2009].

Lionbridge, Standardizing Business Terminology through Localization Management. [Online]. Available at http://www.lionbridge.com/lionbridge/en-US/services/localization-translation/terminology-management.htm [Accessed 3 February 2009).

Warburton, K., 2005, Terminology: Getting Down to Business, *The Globalization Insider* [online]. Available at http://www.lisa.org/globalizationinsider/2005/07/terminology_get.html [accessed 26 January 2006].

Wittner, J., 2007. Unexpected ROI from terminology. *MultiLingual*, (3), p.51-54.