Is “Fuzzy Matching” Still Fuzzy to You?

How translation memory really works

INTRODUCTION

The increasingly stronger demand for translated content has shaped the translation industry. Formerly, very much a craft-like business, it has become a technology-driven industry seeking operational excellence.

The industry has been experimenting with computer aided translation (CAT) tools since the 1980s. Translation Memory, a reusable database of paired text segments in the source and target language, has emerged as the most practical technology and is widely used today.

While for many the translation memory is common and its benefits straightforward, the logic behind it is often puzzling. This paper explains the basics of this technology and what happens behind the scenes.

KEYWORDS: *Computer-Aided Translation, Computer-Assisted Translation, CAT, translation memory, translation technology*

TM versus MT

People often confuse translation memory (TM) with machine translation (MT). So, let’s first get that sorted out. **Machine translation** is done by a computer without any human interaction. Examples of this technology are Systran, Bing Translator, or Google Translate and many are readily available on the Internet free of charge. The resulting translation is usually far from perfect. While useful as a quick translator to understand a meaning of foreign text, the use of machine translations still very limited for commercial use.

People often confuse translation memory (TM) with machine translation (MT).

Translation memory, on the other hand, is generated when a human performs the translation and the computer assists. Translated segments of text, paired with their English originals, are stored in a database and these translated segments are then available for reuse in similar content. Examples of this technology are SDL Trados, Catalyst or WordFast.

Segmentation

Regardless of the original authoring tool, the localization vendor first converts the source files to a pivot format compatible with the translation memory. For example, FrameMaker files are first converted to MIF and then to RTF. Next, the translation memory tool segments the files based on the predefined segmentation rules. Typically, one segment equals one sentence. However, the tool is smart enough to recognize titles, for example, as segments as well.

Translators work directly with the segmented files and translate each text segment one by one. As soon as a segment is translated, its translation is stored in the database and paired with its source language original. It is then available for reuse.

Fuzzy matching

So how does it really work? The translation memory tool analyzes each segment against those stored in the translation memory and looks for matches. If it finds an **exact match** (also known as 100% match), a segment with the same content and formatting, it automatically inserts its translation in the target text. The translator only verifies that it is still valid and that it fits within the context.

Some tools recognize **in-context matches** (also known as 101% matches or perfect matches) where the translator can be certain that the translation is correct because not only the content and the formatting are identical but also the segments appear within the same context.

However, what makes translation memory tools powerful is that they also recognize segments that are not the same but are very similar. These are called **fuzzy matches**. The translator has to make a few changes to the proposed translated text.

Working with tags

Translators work with tagged files and even hidden objects are visible and accessible to translation. In the following example, there are two types of tags: protected tags (grey) and placeable tags (red). The translator cannot alter the protected tags but she can move the placeable tags if the syntax of the translated sentence requires it.

Example:

<p class="title5">Localization is the <i>process of cultural and linguistic adaptation</i> of products and services for foreign locales.</p>

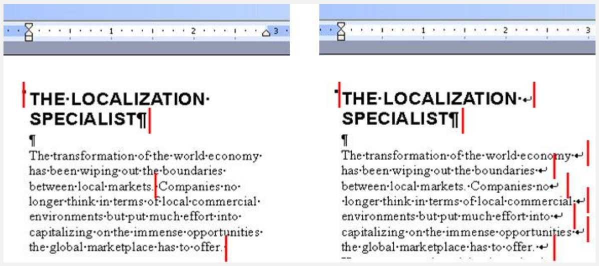
It is important to know that tags and therefore the formatting are stored in the translation memory. This means that two identical sentences may not be matched if the tags are different. The translation memory tool will likely identify them as a fuzzy match and the translator will have to make the formatting adjustments.

Companies sometimes change the formatting of their manuals from one version to another to improve the look and feel or to comply with new corporate image. However, they not always realize the impact this will have on the cost of localization updates. Even if the content changed only minimally, the translation memory returns results indicating that there is only little leverage. While the new formatting may look nicer and be more user friendly, the increased localization cost may not warrant it.

Tags also affect segmentation

As mentioned earlier, translation memory tools have a set of predefined segmentation rules. Based on these rules, translation memory tools break down the text to individual segments. But consider the following example.

Example:



THE LOCALIZATION
SPECIALIST¶
¶
The transformation of the world economy
has been wiping out the boundaries
between local markets. Companies no
longer think in terms of local commercial
environments but put much effort into
capitalizing on the immense opportunities
the global marketplace has to offer.

THE LOCALIZATION ↵
SPECIALIST¶
¶
The transformation of the world economy ↵
has been wiping out the boundaries ↵
between local markets. Companies no↵
longer think in terms of local commercial ↵
environments but put much effort into ↵
capitalizing on the immense opportunities ↵
the global marketplace has to offer. ↵

In the left column, the translation memory tool followed the basic rule that a segment starts with a capital letter and ends with a period. It recognized three segments. However, on the right, the writer had inserted hard-coded carriage returns (<p> tags). The tool used a different segmentation rule where the segment’s beginning and end is marked by a <p> tag and thus it yielded 9 segments! Without a doubt, the translation memory will not find any matches at all.

Suggestions for achieving the highest possible leverage

Finally, here are a few suggestions that will help you achieve the highest possible leverage and thus decrease the cost of your translation projects:

- Understand the segmentation rules and reproduce them from a version to version.
- Avoid formatting changes because formatting tags are stored in the translation memory.
- Standardize text (sentences and even paragraphs) and use it across all of your content as much as practical.
- Avoid wordiness. Translation and localization companies charge per word.

EzGlobe acts as a strategic partner for companies that believe in the importance of addressing their clients, partners or employees in their own language. The company helps its clients go global by providing **professional translation, localization and internationalization services**.

www.ezglobe.com

Sophia Antipolis, France
Tel.: +33 4 92 94 23 90

Massachusetts, USA
Tel.: +1 781 322 0370