MT and TM Technologies in Localization Industry: The Challenge of Integration

Nikolai Puntikov
STAR SPB (Russia) Ltd.
P.O.Box *703.* St.Petersburg 199053
Russian Federation
E-mail: Nick@star.spb.ru

Abstract

The objective of this paper is to clarify certain technological aspects of the localization business process. An introduction to the Translation Memory (TM) technology is provided, followed by an analysis of how TM and Machine Translation (MT), when used together, can increase productivity in software localization workflow applications.

A special section is devoted to the issue of standard exchange mechanisms to represent translation memory data so that they can be shared among users of different TM and MT tools.

1. Introduction: Localization Industry

Software manufacturing is a field where internationalization is such a crucial factor in the success of a product that the localization process often begins while the product is still in its development stage. Activities in the localization business process include:

- Translation of documentation and help files
- Localization of user interfaces
- Verification, validation and software testing
- Maintenance of terminology
- Desktop publishing
- Document management and version control
- Workflow management

A common setup accepted in the localization industry often implies three levels of the business model:

(1) A software publisher is responsible for authoring and preparing language-dependent resources with the aim of making the localization process as quick as possible.

(2) For the purpose of localization, the publisher looks for a service provider to outsource work for as many languages as the provider can handle. In many cases, this also includes compilation and testing of the localized software.

(3) Finally, service providers are no more mere translation agencies. They are international networks that tie together translators and validators spread all over the world.

The Localization Industry Standards Association (LISA) was founded in 1990 in Switzerland as private, non-profit organization. LISA defines its mission as "promoting the localization industry and providing a mechanism and services to enable companies to exchange and share information on the development of processes, tools, technologies and business models connected with localization, internationalization and related topics" ([LISA99]).

LISA's current membership of 130 leading players from all around the world includes software publishers, hardware manufacturers, localization service vendors, and an increasing number of companies from vertical business sectors.

One thing is certain in the world of localization – with the modern requirements to the time-to-market and quality factors, no localization would be possible without appropriate productivity tools.

The focus of this paper is on tools used by service providers for the purpose of translation, verification, validation, and terminology management.

All major developers of commercial tools are LISA members. According to the 1997 and 1998 Localization Industry Surveys ([LISA99]), all localization service providers use either commercial or proprietary tools. In this paper, I will discuss how translation workflow is organized in a typical localization company; define a suitable role for an MT system in this workflow process; and explore what can be done to combine the strong points of MT and another popular technological paradigm called Translation Memory (TM).

2 Translation Memory Technology

The idea of translation memory is basically very simple. It is to save time by prompting the answer to a question which every professional translator inevitably asks: "'How did I translate this sentence the last time I

saw it?". The most basic TM should provide a quick and accurate answer at least to this question. A more sophisticated TM tool is able to:

- automatically translate 100%-matches found in the TM database;
- account for minor differences in the original segment and suggest a translation for a similar segment found in the TM database ("fuzzy" match);
- lookup terminological database(s) for terms in a source segment and display their translations;
- preserve formatting of the original document and restore it after translation is completed.

The keyword behind translation memory technology is "re-use". It does not attempt to replace a qualified human translator. Rather, the goal is to provide the translator with various facilities to increase productivity and to improve translation quality.

Figure 1 shows a typical workflow of a TM tool on the example of the STAR TRANSIT™ technology ([STAR99]).

Different manufacturers follow different technical ideas in implementation of their products. However, the basic components of practically all commercial TM systems are the same. Namely:

- A translation memory database
- An alignment tool
- A terminology management system
- A document editor
- A set of import/export filters

[Benis99] is an excellent reference to a detailed comparative analysis of TM systems offered by 5 different manufacturers: Atril (Déjà Vu), IBM (Translation Manager), SDL (SDLX). STAR (TRANSIT) and TRADOS (Translator's Workbench).

In the following discussion I will attempt to identify a niche in the above workflow for MT systems and NLP research breakthroughs in general. Let's look at the facts: *all* localization service vendors use TM tools and *some also* use MT. The odds are not in favor of a

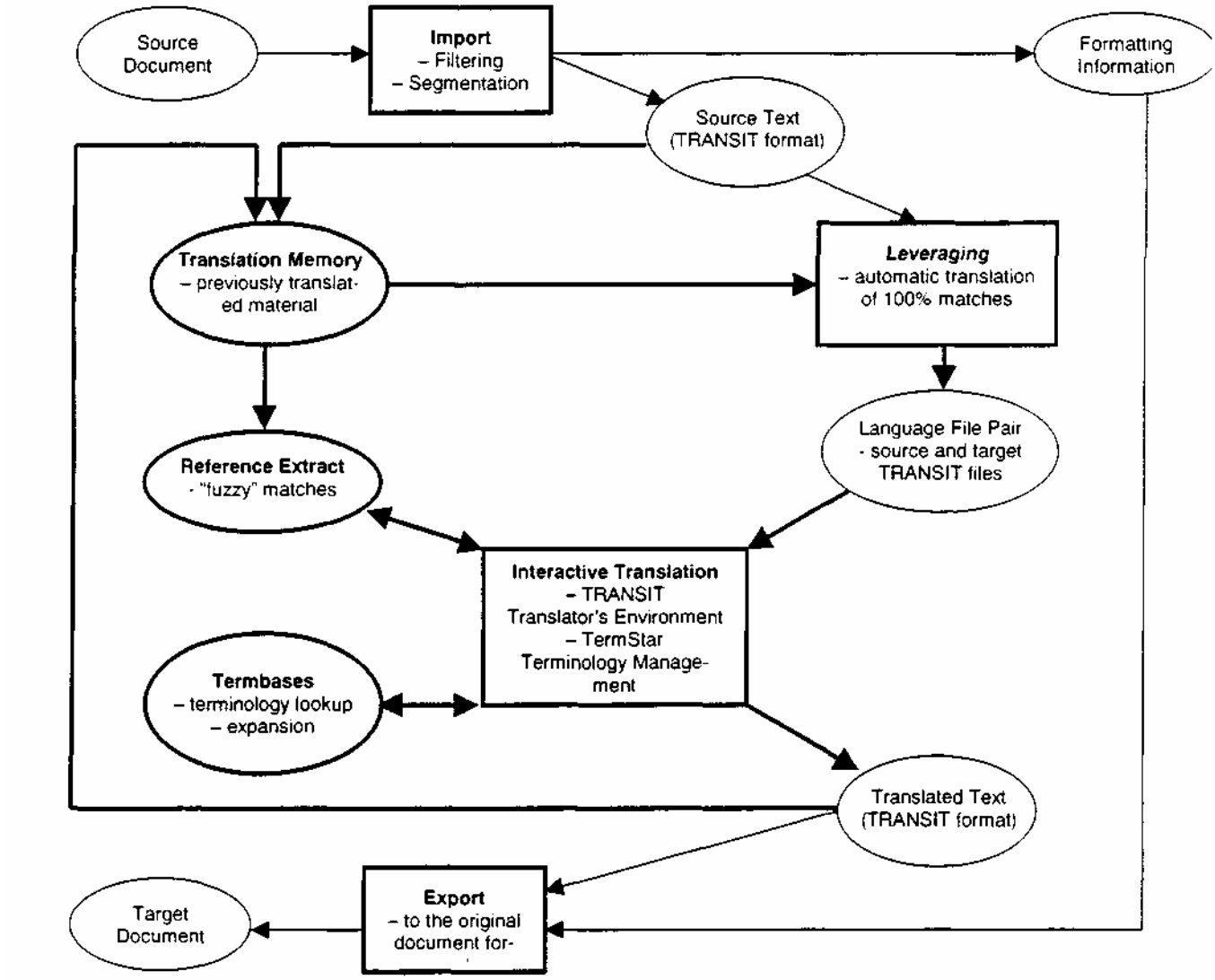


Figure 1. Translating a document with TRANSIT

drastic change in the localization process that would lead to replacement of TM technology with fully automated MT-based processes. No algorithm can ever translate better than a human being.

On the other hand, no human being can be as consistent as a computer program. This is precisely the idea of translation memory: once a text segment is translated and verified, then the next time this or similar segment is encountered it can be translated in exactly the same way.

The next goal is to increase the productivity of translation work to the greatest possible extent. This is where MT can make a significant contribution, especially when a translation project consists of mostly new material (that is, there is not much to leverage from the translation memory).

3 Integrating MT in Localization Workflow

MT and TM technologies are siblings. Like all family they share some things in common:

- Both are computer technologies. That is, they involve the usage of computers for translation.
- Both have to deal with natural language text in a machine-readable form and with lexical resources (lexicons, grammars).
- Both focus on productivity gains in the translation process.

However, as with all siblings, certain differences are obvious. Socially, MT was the favorite daughter, grown up in research laboratories with full governmental support for years before a very first commercial MT system was put on the public market. TM was an ugly duckling shyly moving ahead as a market driven initiative. Technically, the focus of MT is a fully automated process based on fundamental linguistic research. Whereas TM's principal concern is the workflow of the (human) translation effort.

That is probably why many service providers in the localization business only use TM. They have a strong prejudice that post-editing MT output takes more time than translating from scratch.

Companies who use both technologies often do it for historical reasons. Twenty years ago, when TM technology did not exist, they started to work with MT. They hired professional linguists to build specialized lexicons and learned how to customize MT systems for maximum quality output.

At the same time, many of them had in-house software development where the concept of TM had been implemented to some extent. When TM technology matured, these companies had to follow an established trend. One example of such a transition is SAP, a long-term supporter of MT technology. They began introducing TM tools in their workflow applications "in particular for language pairs for which commercial MT systems have not proven efficient enough or are

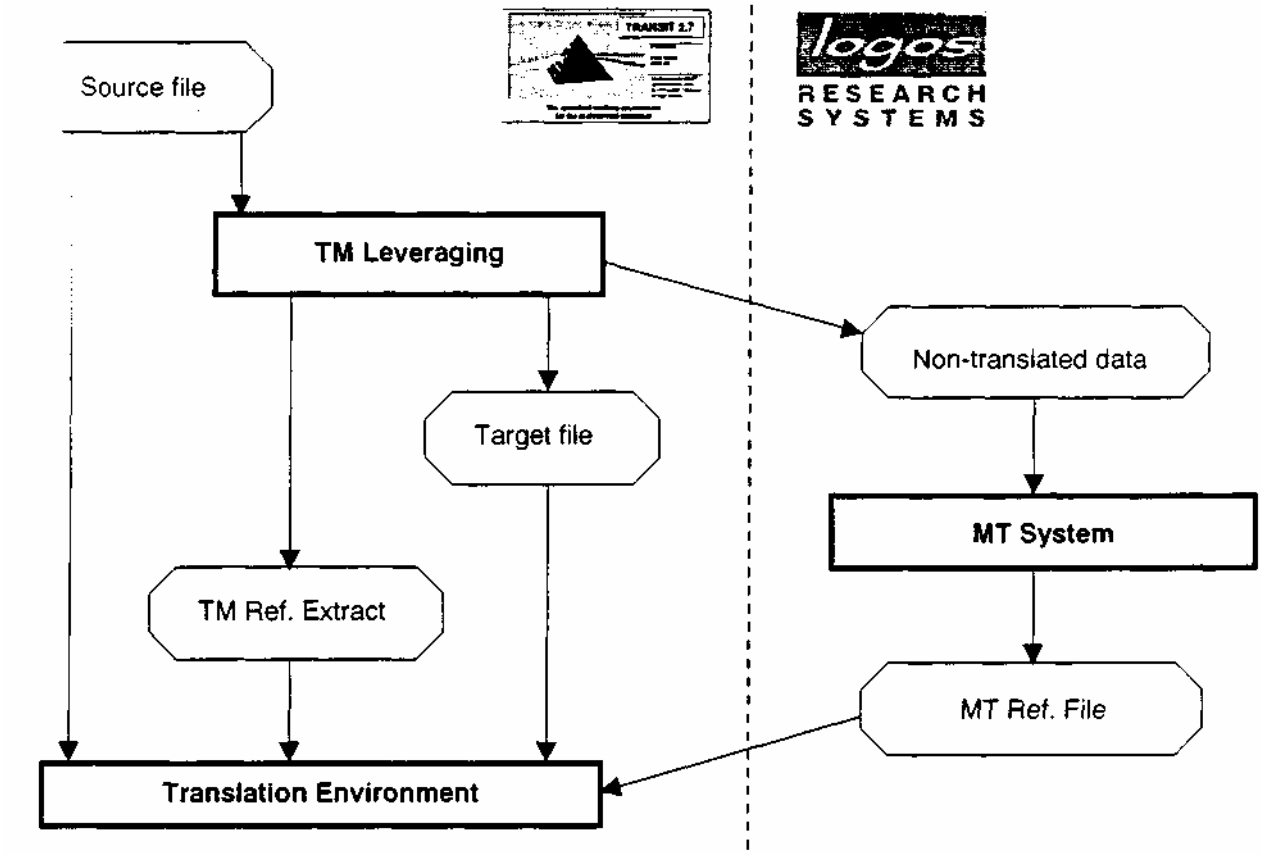


Figure 2. Integration of TRANSIT and LOGOS

not available at all" ([Brundage/McCormick97]).

Figure 2 shows a workflow application based on the LOGOS™ MT and the STAR TRANSIT™ TM technologies. The idea again is simple: whatever was not found in the translation memory database is sent for translation to the MT system. The MT translations will supplement the reference material (highly scored "fuzzy" matches) used by a human translator in the interactive CAT environment. However, the implementation of this approach presents certain technical problems that have to be observed by those planning to introduce the integrated solution in their workflow.

In particular, TM tools normally support a wide range of languages. For the basic product, the only dependence on language that these tools have is limited to character set support. In reality, this is more complicated because many tools use language-specific information for the purpose of segmentation, alignment, leveraging and terminology look-up.

Still, there are no TM tools which would restrict themselves by supporting only one language pair. This is a key difference: MT tools (except for interlingua-based, and those are not yet commercial) support *language pairs,* while TM tools support *languages.* Some tools (like TRANSIT) allow customization, so that users can create their "own languages" by means of defining certain system parameters. Any combination of supported languages is valid in a TM tool.

This difference implies certain restrictions on implementation of integrated applications. A multilingual company would normally use just one TM tool for all their projects. In contrast, many MT tools may be in use at once, in the same company. Often the best MT tool for a given language pair is the one which does not support anything except this specific pair.

Designing a sophisticated API (application programming interface) for all those systems to communicate in a standard way would be a difficult task, if indeed possible. Such a solution may be feasible for a translation group responsible for just one language pair. For example, the translation center of St.Petersburg (Russia) branch of Lucent Technologies integrated TRANSIT and ProMT using an API which communicates with ProMT to translate a segment on demand. However, they only translate from English into Russian. The business justification for a multilingual company is questionable, and, generally, the advantage of this approach is doubtful. If the result of MT translation of a segment is the same, why not do it in a batch mode at the project preparation stage?

A more flexible solution is the one which relies on the exchange of files. That is, a TM system creates a file with segments to be translated and passes it to an MT system. Exactly the same number of translated segments must be returned to the TM system, which will later use them to create its temporary translation memory to be used by a human translator.

As far as TRANSIT is concerned, such an interface is implemented with two MT systems – LOGOS™ and ProMT™. ProMT recently published a press-release announcing similar interface between their MT system and the TRADOS TM system ([ProMT99]).

There are certain difficulties with this approach as well. They concern treatment of formatting markup within segments, context analysis, and other issues to which we will come to later. However, the most interesting problem seems to be a standard format to represent data to be exchanged between various TM and MT systems.

4 TMX: LISA exchange standard for translation memory data

Developers of each TM tool use proprietary formats to store translation memory and terminology data. For example, TRANSIT has no database in a strict sense at all. They advocate a principle of "a single source" and their users create "translation memory" (called "reference material") on-the-fly using language file pairs from previously completed projects.

Taking in account that translation memory collections (databases and file pairs) are considered by localization clients, as well as their service providers, as being capital assets, we have a problem of users being strongly dependent on vendors. Users want the freedom to choose among the components of different systems in order to meet their particular needs. Therefore, it was the objective industry demand that led TM tool manufacturers to understand that support of a standard mechanism for data exchange would actually broaden their potential market.

In about one year the LISA OSCAR[1] SIG (Special Interest Group), comprised of the leading manufacturers of TM tools, developed a fully functional UNICODE-based and XML-compliant standard (TMX) for exchanging translation memories regardless of tool or operating system. I've been honored to be a member of this group from its start. It proved to be a unique example of an extremely fruitful collaboration of competitors for the benefit of the entire industry.

Alan Melby of Brigham Young University, the Technical Secretary of the OSCAR, recently published two papers on this topic. For those interested in OSCAR history and technical features of the LISA standards, I would recommend reading [Carroll99] [Melby99a], [Melby99b] and [LISA99tmx].

Most of the leading TM vendors already announced support for TMX in their tools. Some of them decided to use TMX as their standard import/export format. In

[1] **TMX** stands for **T**ranslation **M**emory e**X**change **OSCAR** (**O**pen **S**tandards for **C**ontainer-and-content **A**llowing **R**e-use) is the LISA Special Interest Group responsible for its definition

is interesting to note that there is not a single commercial MT system, which has taken this step. They have apparently missed the value of a common interchange format. "Some people want to exchange translation memories – what does it have to do with MT?" We shall see whether this is a valid point.

The file exchange approach to interoperation between TM and MT tools is discussed above. As an MT vendor, one cannot anticipate (and follow up) all formats used by different TM tools. As a TM vendor, one does not like to isolate a segment from its embedded formatting. Rather, it is highly desirable for the MT system to move tags occurring in a source segment to their respective semantic positions in the translation. As a result, each and every pair of potential collaborators (TM+MT) has to meet face to face and discuss details of their interface even if it is as simple as the exchange of files.

TMX offers a solution to this problem; see Figure 3. If a TM tool already supports TMX, exporting non-translated segments in the TMX format and importing back translated segments will be almost no work. If an MT tool is able to read and interpret TMX, it can easily be integrated into any workflow application. System developers no longer need to be concerned about which particular TM provides source data or which particular MT translates them.

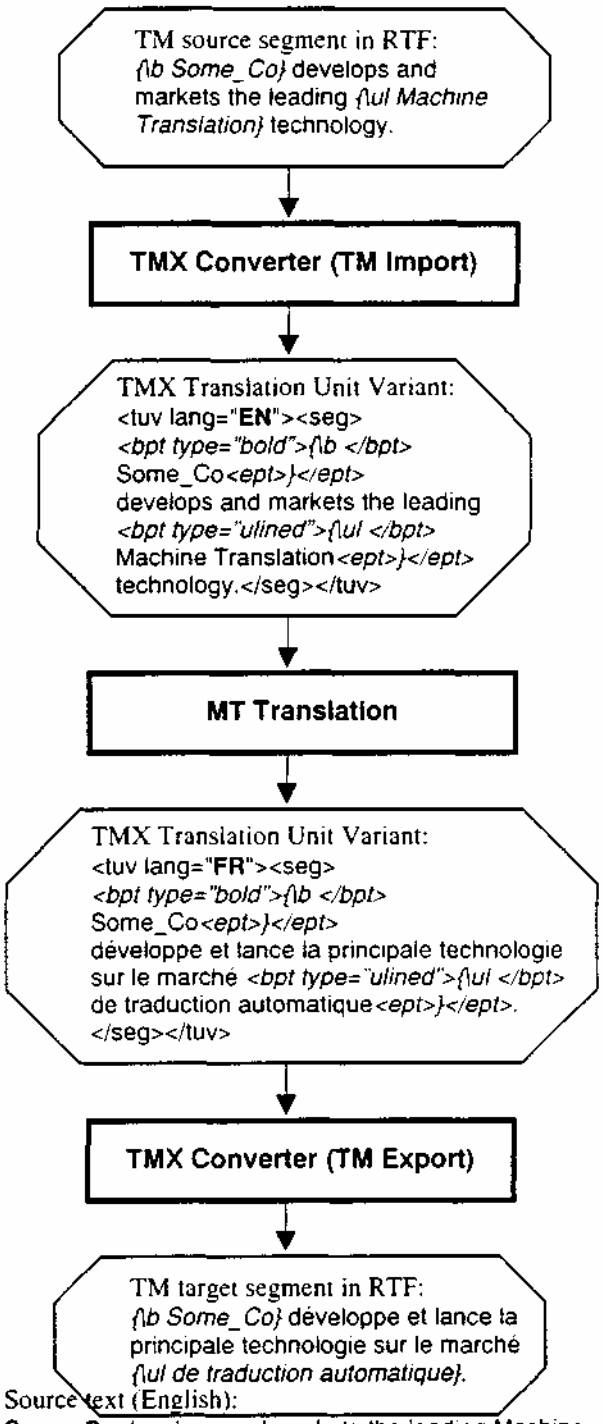
Another very important benefit is that through TMX, a TM tool will reveal to an MT system just enough semantics of the inline tags. Figure 3 shows an example of a TMX segment. All tags within a segment are encapsulated using XML markup. As a minimum, this markup supports recognition and separation of the text from the tags. That is, an MT system which does not care about markup has an easy way to skip all formatting information.

But there is more; the TMX content markup provides information on how the tags are related to the text. It defines:

- paired tags which surround text fragments (like "bold")
- isolated tags (whose ending/beginning counterpart is not found in the segment)
- placeholders (with an optional attribute "assoc" used to define whether a placeholder is associated with the previous or the following text)
- subflows (for example, the definition of a footnote or the text of a title in a HTML anchor element).

Optionally, a TMX file vendor may provide a "type" attribute for each content tag. The types define exact meaning of the tags (e.g. "bold", "bookmark", "footnote"). However, for the purpose of machine translation this information seems to be already unnecessary.

TMX is not the first tagging format for parallel texts (see, for example. [Thurmair97]). Its importance lies in the fact that TMX is the first initiative implemented



Source text (English):
Some_Co develops and markets the leading Machine Translation technology.

Target text (French):
Some_Co développe et lance la principale technologie sur le marché de traduction automatique.

Figure 3. MT&TM Interoperation with TMX

by TM tool developers. In record short terms it was accepted by the localization industry as their de-facto standard for the exchange of translation memories. It is probably only a matter of time for the MT system manufactures to take commercial advantage of this interchange format.

It should be noted that supporting TMX is an easy matter for MT developers. Most of their systems already know how to parse HTML files since translation on the web is one of the hottest applications of commercial tools. TMX is an XML application. XML, in turn, will soon become a successor to HTML as the Internet standard markup language.

5 MT&TM: Other Challenges of Interoperation

Beyond the usual concern about the quality of machine translation, there are several other problems which localization service providers observe when they think of using MT systems in their workflow applications.

5.1 Formatting Information

A strict distinction between translating and DTP work is a thing of the past. Today, translators are responsible for providing their work in a format, which matches that of the original document. However, localization companies cannot expect a translator to be familiar with all DTP packages the clients may use. That is why filter software has become a very important part of commercial TM systems. The filters convert original documents into a format which is used in the translation system environment. TRANSIT, for example, provides its own editor and uses a plain text format with simple tags, while TRADOS converts to an extended RTF so that translators work in WinWord. After translation is completed the filters merge translated text and formatting codes to rebuild the target document formatted as the original.

The importing program leaves certain inline tags in the source segments. Translators know the meaning of these tags and are instructed to keep them in the correct place in the translated segments.

Do MT systems perform this function? Not always. For example, in the TRANSIT&ProMT workflow application all TRANSIT tags are encapsulated with an additional markup, so that ProMT distinguishes them from the text. This is similar to what TMX does, but weaker, because no semantic information is available with the tags. The TM system just places all tags at the end of the translated segment. Moving them to their correct position is a manual process. As a result, many segments have to be post-edited even if they were translated linguistically perfect.

If an MT system does not solve formatting issues, there is probably no worthwhile advice. Such a system will most likely fail on the localization market.

If a good system implements an algorithm to respect tags, then it needs to know what the tags mean. As in the above example (Figure 3). knowing that a pair of tags “begin bold”/ “end bold” surrounds a piece of text would allow the MT system to place those tags in the target segment exactly where they have to be.

But there is another obstacle. There are many common text formatting schemes. TM tools add to this their proprietary formats. MT systems simply cannot support all of these at the same time. There are several initiatives to create a universal scheme for the text markup ([Thurmair97]). However, there is also the understanding that none of these schemes can fully cover the dynamic and ever-changing clients’ needs. This is the underlying power beneath SGML and its subset XML, which are meta-languages to create particular markup languages. It is probably this approach which has a future.

The current version 1.1 of TMX is generic enough to allow a universal means of data exchange between TM and MT. If some changes to the standard are justified to enhance this, the OSCAR Steering Committee invites proposals for the subsequent versions of the standard.

5.2 Robustness of MT systems

It often happens that MT systems would work well on grammatically correct input supported with a full lexicon, and would be useless when the input contains partial or broken sentences. As a matter of fact, application of TM technology in some sense adds to this problem.

The TM tools deal with segments. They have segmentation rules which may be hard-coded or controlled by means of regular expressions. Those rules respect sentence boundaries, but they also respect document layout (markup). Therefore, sometimes a segment will contain a few words which form no sentence (consider a `title` or a cell in a table).

Another problem occurs when the approach *translate only non-translated* is followed. The TM extract file submitted for translation contains individual segments pulled out of context. In this case, sophisticated MT tools will not be able to apply their algorithms for context analysis.

Last (but not least), given tight deadlines of localization projects, a user may not have enough time to complete the MT lexicon with new terminology. In the best case, bilingual term pairs will be provided lacking features required for MT processing.

The more robust the MT system, therefore having the ability to produce meaningful results with insufficient information, the better are its chances on the localization market.

On the other hand, cooperating with TM tool developers can also be expected to give advantages. For example, to deal with discourse problem, one possibility is to supply the complete text in which non-

translated segments are marked up in a way understood by the MT system. In fact, if the MT component would care to take in account verified translations already available in pre-translated segments of such a file, it may produce even better results.

5.3 Lexicon acquisition

There is a commonly accepted process in localization workflow for expanding a human-oriented terminology database. In most cases, translators are responsible for adding pairs of terms and relevant context examples as they work. At the completion of a translation project professional terminologists and technical experts would validate and approve the new terminology.

The fact that building an MT lexicon requires special qualification (which implies a mysterious knowledge of formal linguistics) intimidates potential users. A small agency probably cannot afford to hire a professional linguist. Even with companies that can afford a full-time specialist, replicating the terminology in two lexicons would be a big cost issue. Besides, introducing a new process in the well-proven workflow application is not an easy thing to do.

A promising solution would probably be a system with both TM and MT implemented in one shell. An example of this type of system is T1, which grew from the METAL MT system ([Schwall/Thurmair97]). However, there is a need to support languages beyond those that its MT component supports. If T1 were as powerful in supporting multiple language projects, multilingual conceptual lexicons, and multiple document formatting schemes, as the best TM tools on the market, it would have a great potential on the localization market. In the meantime, our interest is in applications based on a single powerful TM technology and third-party MT systems dealing with different language pairs.

The problem of maintaining MT lexicons would be easier if MT systems provided fully automatic tools to construct their lexicons. There are several research paradigms in this area.

One is an old challenge of NLP to acquire bilingual lexicons from aligned parallel corpus. The methods include noun phrase identification, discourse analysis, statistical inference, and so on. Apparently, a particular translation memory database is the aligned corpus.

The other approach, which comes from developers of commercial MT systems, is to provide lexicographers with semi-automated "fast coding" methods based on bilingual concordances. A good MT system offers a combination of these and other methods. See, for example, [Gerber/Yang97] where SYSTRAN dictionary development is presented.

Finally, a lot of work is done on implementing standard mechanisms for terminology exchange. For example, the MARTIF ISO standard targets on exchange of termbases, while the OLIF format defined within

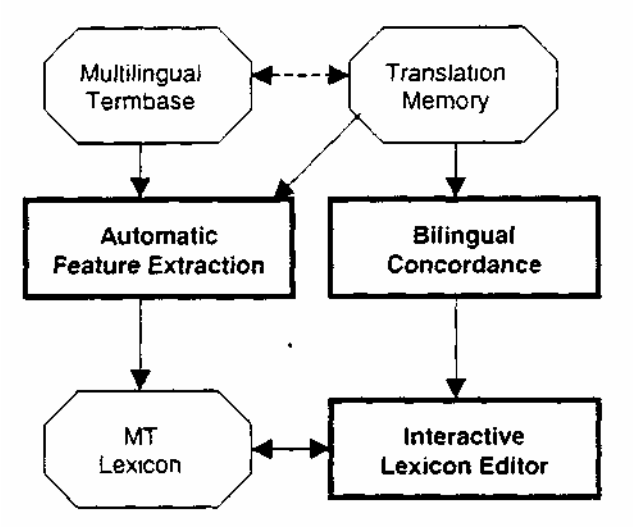


Figure 4. Automated construction of MT lexicon

the OTELO project supports multifunctional terminological databases *(lex/term-bases,* see [Melby/Wright99]). LISA's OSCAR also pays attention to this matter. Its current project is a standard for TermBase eXchange (TBX) in the localization industry.

Figure 4 suggests a process in which the above techniques are integrated for the purpose of semi-automated construction of MT lexicons.

Translation memory and termbase when used together, provide the basic linguistic information which could be automatically used to assign features of an entry in the MT lexicon. Such as:

- Morphological class (one word ending or article may be enough to infer the entire class)
- Syntactic pattern
- Co-occurrence information

Yet another suggestion would be to account for the additional information which TM offers. Both structural and inline formatting tags are normally available in translation memory and they are appropriately aligned. Many heuristics could be developed which would make use of such information. Let's take two simple examples:

- If a polysemous term in a cell of a table within a given project was translated in a certain way, chances are good that the next time it occurs in another cell or table, it is to be translated in the same way.
- If a noun phrase in the source segment is highlighted with font attributes, it is a candidate for a multiword term. Its translation is probably a text fragment in the target segment with similar highlighting. The TM can be used for the statistical verification of this hypothesis, and for the automatic extraction of the agreement models.

It is not common today to use such information for the MT lexicon construction, but it clearly makes sense to research potential benefits of these heuristics. Briefly, it is about using two sources of information (not only one of them) when transferring terminology from a human-oriented termbase to an MT Lexicon: terms defined in the termbase; the translation memory, in which these terms are used.

6 Conclusion

Integration of MT and TM technologies is a multi-factor process which requires efforts from representatives of both paradigms. In this paper, I have addressed issues concerning interoperation on a higher level. When complete and self-sufficient MT and TM systems supplement each other in a workflow application, this can potentially increase translation productivity.

There are other interesting opportunities which exist on a lower level of cooperation. This refers to both implementation of NLP algorithms in TM systems and introducing TM-based techniques in MT systems. As noted in [Thurmair97] "the boundary between terms and (memory) phrases and between memories and (multilingual) texts become less and less clear".

Certain work in this direction goes on in both communities. On the localization business side, consider, for example. TRANSIT's alignment tool which is a highly automated program using all linguistic, statistical and formatting information available in a document and in the TRANSIT's entire translation memory.

TM developers are interested in advanced algorithms for segmentation (in particular, phrase extraction going beyond sentence-based segmentation), morphological analysis including multi-word terminology processing, fuzzy matching, term substitution, and more.

The future of the localization process is in components which could be merged together to resolve the client's problems in a most efficient way. These components are required for many languages as modules which could be easily plugged into current TM-based applications. That is what the localization industry expects these days from the TM research and development community.

As far as the integration of MT systems into localization workflow is concerned, there is a strong promise that close contacts between developers of both technologies will help to resolve problems observed in this paper, as well as other problems preventing wider usage of MT in localization business applications. So far, LISA has proved to be a reliable and inspiring umbrella for this information exchange.

References

Benis M. (1999). "Translation Memory from O1 to R1". ITI Bulletin, April 1999. UK Institute of Translation & Interpreting.

Brundage J., McCormick S. (1997). "Managing Distributed MT Projects Today: A New Challenge", proceedings of MT Summit VI.

Carroll S. (1999). "The Technical Content of TMX 1.1: A Format for the exchange of data between competing translation database systems". Multilingual Computing & Technology. No.22. vol.9/6.

Gerber L., Yang J. (1997). "SYSTRAN MT Dictionary Development". In proceedings of MT Summit V!

LISA99: Localization Industry Standards Association. http://www.lisa.org

LISA99tmx: Translation Memory Exchange. http://www.lisa.org/tmx

Melby A. (1999a). "TMX has been Implemented!" LISA Newsletter, vol.VIII. No.3, September 1999.

Melby A. (1999b). "Sharing of Translation Memory Databases Derived from Parallel Text". Forthcoming in the Kluwer book on parallel text processing.

Melby A., Wright S.E. (1999). "Leveraging Terminological Data for Use in Conjunction with Lexcographical Resources". Forthcoming. Available - http://www.ttt.org

ProMT99: http://www.promt.ru

Schwall U., Thurmair G. (1997). "From METAL T1: Systems and Components for Machine Translation Applications". In proceedings of MT Summit VI.

STAR99: http://www.star-ag.ch

Thurmair G. (1997). "Exchange Interfaces for Translation Tools". In proceedings of MT Summit VI.