MACHINE TRANSLATION
- EVOLUTION NOT REVOLUTION -

Jennifer A. Brundage

SAP AG
P.O. Box 14 61, D-69185 Walldorf
Germany
jennifer.brundage@sap.com

Abstract

The continuous trend towards globalization means that even the most modern of industries must constantly re-evaluate its strategies and adapt to new technologies. This not only involves living up to the demands set by the product life cycles but also to find solutions satisfying additional internal needs. As a long-time supporter of MT and TM technology, SAP has shown that it can make productive use of competitive, commercial NLP products. As a first step, an integrated solution using TM together with MT was targeted. Having implemented different solutions for two types of documentation, the focus is now on not merely to integrate other technologies (e.g. terminology mining or controlled language) but to provide a uniform solution for processing any type of text. This involves not only supporting the needs of technical writers and translators but of all employees in their multilingual working environment.

Keywords

Machine Translation, MT evaluation, MT entry generation, NLP tools integration, workflow tool

NLP Technology in Documentation and Translation at SAP

From the late 80’s early 90’s, SAP has developed and enhanced its own proprietary terminology database (SAPterm) which is fully integrated in their business software product. Currently, version 3 is being used productively. In-house it is the central source for SAP terminology and can be accessed by all SAP employees via a Web interface.

The MultiLingual Technology (MLT) group originally started out as a purely Machine Translation group back in 1990. Once its first project to introduce MT for German-to-English as an in-house service for the specialized translators was realized, the group very soon realized that more could be achieved with such technology. The success of this initial project then triggered several others.

With the importance of the SAP products in the market growing, the number of customer notes[1] (or short: notes) increased accordingly. This posed a problem since the translation had to be available within 24 hours upon release of the source text for the priority 1 notes (Support Level Agreement). These parameters made it impossible to succeed using the conventional methods. In 1994, a cooperation with the MT vendor to implement a workflow tool combining Machine Translation and Translation Memory for the translation of notes was therefore initiated. With this tool, it was possible for 4-6 people to easily translate some 100 to 150 notes (on average 96 words) per day. Once the entire process was set up, thoroughly tested and the technology proved to be stable, the customer notes translation was gradually outsourced starting in 1996 (Brundage, McCormick & Pyne 1997).

SAP’s popularity in other than the English and German speaking markets, of course, led to the support of other languages having to be extended. Thus, Machine Translation for English-to-French was introduced in 1994 as well. The user lexicon currently contains some 60 000 entries. Since none of the commercial MT systems supports all languages and large part of the documentation was being re-used, alternatives needed to be found. In 1996, standard commercial Translation Memory technology was introduced to SAP on a large scale. As the demand cropped up, MT for English-to-Spanish was set up in 1999. To shorten ramp-up time, a new method was used. All English source words maintained in SAPterm without a Spanish target equivalent were exported and existing translation memories for that language pair were searched for possible translations. The proposals found in the memories were passed on to the specialist translators together with the context information. These then verified or changed the proposals. The result was a list of bilingual term pairs. This list could then be imported into the MT system (overall time saving: 20%) and also serve as a basis for completing the entries in SAPterm. The MT user lexicon now contains some 58 000 entries.

The MT system used for German-to-English not being Y2K-compliant represented a new challenge for the group. Not only had the follow-up MT system to be evaluated and the terminology migrated, but a completely new workflow tool for the notes translation had to be designed, programmed, tested and implemented (Wells 2001). This also involved a rollout plan for the agency translating the customer notes. One huge advantage of this new solution is its user friendliness. The translator controls the entire process via a GUI (see Fig. 1) instead of Unix commands and can use MS Word, which is automatically launched, as editor instead of emacs.

[1] As one of the world’s largest software houses, SAP AG has a complex system for informing customers about bug fixes in and distributing minor source code corrections for the R/3 System, as well as for distributing additional functionality, best practices and other customer requests.

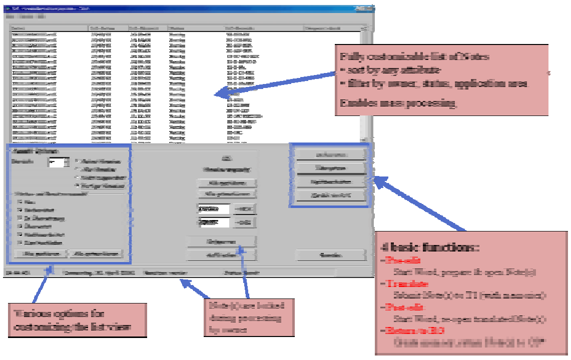

Fully customizable list of Notes
• sort by any attribute
• filter by owner, status, application area
Enables mass processing
4 basic functions:
-Pre-edit
Start Word, prepare & open Note(s)
-Translate
Submit Note(s) to T1 (with memories)
-Post-edit
Start Word, re-open translated Note(s)
-Return to R3
Various options for customizing the list view
Note(s) are locked during processing by owner

Figure 1: End-User GUI for Managing Notes Translation

Also, the administrator can easily configure the application so that the agency is provided with the required files from the various locations. A logging mechanism makes it possible to monitor the process real-time or at a later time.

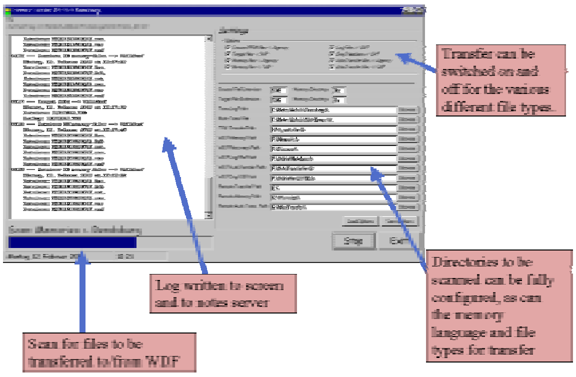

Transfer can be switched on and off for the various different file types.
Directories to be scanned can be fully configured, as can the memory language and file types for transfer
Log written to screen and to notes server
Scan for files to be transferred to/from WDF
Stop
Exit

Figure 2: Interface for Administrating the File Transfer

Currently, two agencies at three locations are dealing with the notes translation. From the start of this project to this very day, the volume of notes being released each day has increased to 450-550 (approx. 180 new and 350 updates). The MT user lexicon for German-to-English contains some 75 000 entries.

New Responsibilities

With SAP growing at the speed it has over the last 5 years, one of the company's current focuses is to reduce costs, but at the same time ensure quality and time-to-market shipping. One means to achieve this is to introduce tools for certain tasks that can be automated. Currently, SAP supports 30+ languages. And with the importance of the market and the multiplication of products offered, the volume of documentation increases. For those languages not supported by the commercial MT systems, a translation memory system is used together with automatic terminology recognition. This solution proved effective and led to savings of up to 40%. Where Machine Translation could be used together with TM, savings reached 60%. However, SAP not only concentrates on using translation tools but also other NLP tools. The volume of documentation SAP has to deal with in the various languages makes it difficult to manage. Are there pieces of documentation that can be re-used? How to ensure terminology consistency? How to enforce the defined standards and guidelines?

The fact that the development of the terminology database SAPterm and of the proprietary translation tools are now part of the MLT group is a great advantage for tackling these issues.

As generally known, translation quality - not only MT output – depends heavily on the quality of the source text. So if SAP succeeded in improving the quality of their source documentation (currently German and English), the costs for the follow-up corrections could be reduced dramatically.

In addition, if unknown or new terms could be identified, defined and maintained before the documentation is released for translation, this would reduce the number of inquiries by the agency; resulting in not only money being saved but also time - another factor not to be underestimated. Ideally, the terminology should be maintained only once and then be available for all NLP tools requiring this kind of resources. The challenge is not only that the various NLP tools need a different set of features but they also require these being available in different formats for usage in different databases.

To top things off, the different document types are written in different formats. A big chunk of the SAP system documentation (such as customer notes, release notes, F1 help) is written in ITF, a proprietary format, for which a converter to and from RTF has been written, thus making the usage of NLP tools easier. Other formats used are the standard Word documents, PowerPoint export files, and HTML/XML.

Ensuring Documentation Quality

Two approaches were taken at SAP to tackle this problem.

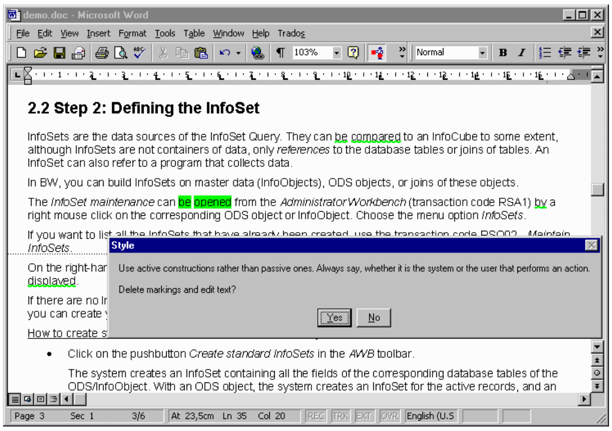
Copy Editing

Copy editing being a traditional "human" approach, it was the most obvious first step to ensure the adherence to SAP's welldefined Standards and Guidelines. Quite quickly, however, it revealed some considerable disadvantages. Only small portions could be dealt with in the given time frame. Being a manual process, it proved to be too time-consuming and too costly.

The SKATE project

As an alternative and extension, the introduction of Controlled Language was discussed. As a result of this, the development phase of the SKATE project was initiated in 2000 with the support of the Board of Directors**.** MLT has always been of the opinion – and still is – that it is not necessary to use Controlled Language to make MT work. Without doubt, however, it can improve the output quality. SKATE stands for SAP Knowledge Authoring – Text Enhancement and is the project dealing with the introduction of Controlled Language tools at SAP. The main objective of this project is to reduce ambiguity, redundancy, size, and complexity of the German and English texts since these are the source for all other languages. As a result, the texts should be more consistent, easier to understand and therefore to translate. Again leading to reduction of costs for the translation into up to 30+ languages.

Having evaluated the market in 1999-2000, SAP decided not to introduce an off-the-shelf product, but rather set up a project using an existing project prototype, in which their requirements could be considered. The advantages of this semi-automatic solution are many. Firstly, the tool can be used by the authors themselves making them aware of the stylistic, grammatical, spelling, and terminological issues of the language they are writing in. In addition, they have the defined standards and guidelines at their fingertips and do not have to look up the various rules themselves. In addition, it is integrated into MS Word (see Fig. 3), the standard editor at SAP. Having a tool to support them for the entire process, larger volumes with a certain quality level can be released in a set time frame.


demo.doc - Microsoft Word
File Edit View Insert Format Tools Table Window Help Trados
103%
Normal
B I
Page 3 Sec 1 3/6 At 23,5cm Ln 35 Col 20 REC TRK EXT OVR English (U.S
Style
Yes
No

2.2 Step 2: Defining the InfoSet

InfoSets are the data sources of the InfoSet Query. They can be compared to an InfoCube to some extent, although InfoSets are not containers of data, only *references* to the database tables or joins of tables. An InfoSet can also refer to a program that collects data.

In BW, you can build InfoSets on master data (InfoObjects), ODS objects, or joins of these objects.

The *InfoSet maintenance* can be opened from the *Administrator Workbench* (transaction code RSA1) by a right mouse click on the corresponding ODS object or InfoObject. Choose the menu option *InfoSets*.

If you want to list all the InfoSets that have already been created, use the transaction code RSQ02 – *Maintain InfoSets*.

On the right-har
displayed.

If there are no Ir
you can create y

How to create s

- Click on the pushbutton *Create standard InfoSets* in the *AWB* toolbar.

  The system creates an InfoSet containing all the fields of the corresponding database tables of the ODS/InfoObject. With an ODS object, the system creates an InfoSet for the active records, and an

Use active constructions rather than passive ones. Always say, whether it is the system or the user that performs an action.

Delete markings and edit text?

Figure 3: The Skate Tool at Use

SAP is planning to go productive this year for both German and English. This means, that not only authors writing either in English or in German can benefit from this technology, but also translators when translating into German or English. A third type of customer for this tool could be a manager of translation project, for example. Using this tool allows logging certain information on the text being checked. Therefore, useful information can be quickly provided for the project manager performing a goods receipt quality check when the translation is returned from an agency or when mass data has to be checked just before release deadline.

SAP has an extra-ordinary plan - running the SKATE tool on raw MT output, MLT hopes to find out how post-editing could benefit from this language checking technology.

Terminology Management

Companies are more and more realizing that terminology is an important investment. Consistent terminology is an important factor when it comes to acceptance by the customer.

Defining and maintaining terminology is a time-consuming and costly task. More so, when you are using several tools requiring different formats.

Standard Exchange Format - OLIF

The initiative to define a standard format for the exchange of terminology between NLP tools – in particular MT systems – has its origin in the EU-funded project OTELO, from which a first version resulted. Soon after the OTELO project ended, the OLIF consortium, headed by SAP, was founded. A new XML-compliant version of OLIF, has become available in April 2001. For more information on OLIF, please refer to www.olif.net and Lieske, McCormick & Thurmair. (2001).

Terminology Mining

Another challenge SAP constantly faces is the development of new industry solutions or add-ons to their standard business software. This of course involves identifying and defining new terms. One means to catch the terminology that slipped through when the documentation was written is to use a Terminology Mining tool. These tools are designed to analyze the text sentence-by-sentence and propose potential term candidates, either single words or multiword expressions. SAP started a project for both German and English in 1999/2000. Having completed the prototype testing, development is now under way. The goal is to go productive for the two source languages this year. SAP is using a commercial tool to identify potential candidates but has stuck some filters on top to improve the results. These filters, for example, identify duplicates in form of inflectional variants and reduces the proposals to one, or eliminates multi-words containing adjectives belonging to the general vocabulary of the respective language.

(Semi-)Automatically Generating MT Entries

This refers to a component being developed in the framework of the EU-funded project TQPro (see www.tqpro.de for more information). The objective here is to fill up potential term candidates (e.g. identified by a term mining component) with lexical information needed by MT systems. The list of bilingual terms in OLIF format is then automatically compared against the MT lexicon to filter out the really new terminology. The challenge here is that the various MT systems treat context information differently or not at all. Can an automatic compare be 100%? How are identical entries that differ only with regards to subject area to be treated? The answer is probably, No. The compromise one will have to make – at least, in a first version- is that only basic features are checked. The final bilingual term list is then imported into the MT system in question. The recommendation is to check the imported list before permanently storing them in the MT lexicon.

Other terminology-related approaches are also taken. For example, a terminology checking tool that can be used as part of the SKATE project or when checking the contents of translation memories and target documents returned from agencies.

The Flexibility/Integration Capability of Commercial NLP Tools

In the past, the NLP tool vendors – particularly the MT vendors – were very reluctant about opening up their systems for integration into environments other than their own. With the prospects the popularity of Web opened up, the NLP vendors suddenly realized that integration could indeed be attractive.

It were the TM vendors who made a first – and very successful – attempt in opening up their systems for integration into customer environments, providing a detailed description of their APIs.

Various projects – cooperation between customer and vendor or EU-funded – were initiated. At SAP, the workflow tool for the notes translation is to be mentioned here, where the MT vendor provided a dll interface for integration. Also, in the EU-funded project OTELO, open API was an issue, which is further pursued in TQPro.

Regarding the support of OLIF, the major tool providers made a commitment to support it. This includes Sail Labs, Logos, Systran, Trados, Xerox, and many others.

Facing the New Challenge

This paper might make the impression of being simply a conglomeration of subjects and aspects, but all these topics are a piece of the puzzle and make up every day work of the MLT group. The group not only concerns itself with contents-related issues (linguistic, lexicographical, quality-related) but also with technical issues (customizability, scalability, integration capability). Every one of them is very important from a user's point of view like SAP, where a large range of languages has to be dealt with. The MT systems available on the market not all support the same document formats or all languages. Even where the same language pairs are covered, there can be considerable differences with regards to output quality, terminology maintenance, and customizability. So, there should be the opportunity to choose the best system for the set requirements. Open APIs and OLIF as terminology exchange format will be a definite plus, especially when integrating an MT system into a workflow tool containing a terminology mining component, for example. The potential term candidates being available in OLIF format can then be directly imported into the MT system.

The MLT group has recently been asked to set up Machine Translation for English-to-Japanese and Japanese-to-English to support with professional translation as well as with the communication for call centers (customer and internal messages and notes, etc.). Triggered by this opportunity, MLT is hoping to realize an all-comprising, platform-independent authoring and translation environment, regardless of what document type is to be processed in which language and for what purpose. All tools are run and maintained centrally, the users having remote access to whatever services they want to use. The front-end user interface of the workflow tool should be customizable to fit any user profile.

Various scenarios are possible. A technical writer could make use of the SKATE tool, the terminology mining component, and could include a dictionary lookup function in his view. In cases, where he re-uses text objects from previous versions and some terms have been changed in the meantime, he could also activate an Intelligent Find and Replace tool. A translator, on the other hand, could use TM and MT technology together with terminology mining. Since SKATE will be available for both German and English, it can also be used in the translation process, when translating from German into English, or vice versa. Other components developed within TQPro, like translatability (Underwood, Jongejan 2001) or completeness check, could also be activated upon request.

As additional language pairs need to be supported, new MT system might be added. And where no MT system is available or does not meet the requirements regarding output quality, customizability, or integration, the huge volumes of bilingual segments stored in SAP's translation memories could serve as a valuable input for an example-based approach to MT.

When this paper was submitted, the MLT group was in the process of evaluating commercially available MT systems. The purpose was not only to update the group's knowledge regarding this type of technology, but also to evaluate the output quality for language pairs to be introduced at SAP. In addition, the group might be forced to replace the MT technology used for some of the language pairs. A pragmatic approach was chosen for the evaluation. Apart from some functional and technical aspects, the first main focus is on the output quality. A corpus containing different documentation types (user manuals, official mails, customer notes, messages, etc.) was set up for each source language involved. These texts were then translated using the various MT systems as delivered, i.e. only using the system dictionaries. The MT output was then viewed without reference to the source text and rated. The profile of the people viewing and rating the output ranges from MT specialists to persons unfamiliar with MT output. The grades that were used are the following:

1 – Good translation
2 – OK
3 – Understandable
4 – Understandable from context only
5 – Not usable

Then the MT output was closely searched for terms/phrases to be maintained in the respective MT system. Once the terminology maintenance was completed, the texts were retranslated and the output evaluated using the same rating system.

One very important aspect to be considered in case an MT system has to be replaced for an already productive language pair, is the size, the quality, and the coverage of the existing user dictionary. And in this context – how easily can the existing dictionaries be transferred to the new MT system (if at all) and how much information will be lost.

The results of this evaluation and the experiences made within will be presented at the MT Summit.

References

[1] Brundage, J., McCormick, S., Pyne, Chr. 1997: Managing Distributed MT Projects Today – A New Challenge, MT Summit VI Proceedings; pp.58-63.

[2] Lieske, C., McCormick, S., Thurmair, G. 2001: The Open Lexicon Interchange Format (OLIF) Comes of Age, MT Summit VIII Proceedings, pp.?

[3] OLIF homepage: www.olif.net

[4] TQPro homepage: www.tqpro.de

[5] Underwood, N., Jongejan, B. 2001: Translatability Checker: a tool to support the decision on whether to use MT, MT Summit VIII Proceedings, pp.?

[6] Wells, J.: Notes Translation Using MT and TM???, not yet published.