How to implement machine translation

Uwe Muegge
How to Implement Machine Translation

Overview

Introduction
This document illustrates a unique approach to machine translation.

The first part of this document provides a brief introduction to machine translation technology. The second part discusses some of the challenges of machine translation. The third part highlights some of the features that set the author’s approach apart from that of our competitors. The fourth part presents an extremely efficient but also complex implementation scenario for machine translation. The fifth part describes a less complex but still highly efficient implementation scenario for machine translation. The last part provides a brief overview of the author’s qualifications.

Contents
Definitions............................................................................................................1
Machine Translation Basics..................................................................................3
Five Machine Translation Challenges.................................................................5
Five Machine Translation Solutions.....................................................................7
Implementation Scenario 1..................................................................................12
Implementation Scenario 2..................................................................................14
About the Author.................................................................................................16
## Definitions

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
</table>
| Controlled language           | A controlled language is a natural language. This is a language where comprehensive rules restrict the use of terms and sentence structures more than in the general language.  
**NOTE:** Machine translation and translation memory systems are particularly effective in combination with a controlled language. |
| Dictionary                    | A dictionary is a terminological collection. This is a collection of dictionary entries in one or more languages.                                                                                           |
| Dictionary entry              | A dictionary entry is a terminological unit. This is the unit that allows translation professionals to store the information relating to a concept.                                                        |
| Machine translation           | Machine translation is a linguistic process. This is the process of translating a source text into a target text where a computer program performs the actual translation. 
**NOTE:** Machine translation is particularly effective in combination with a controlled language and a translation-memory system. |
| Post-editing                  | Post-editing is a linguistic process. This is the process of reworking a machine translation for content and style by a translation professional. 
**NOTE:** Post-editing is particularly effective in a translation-memory system.                                                                 |
| Source language               | A source language is a natural language. This is the language that translation professionals translate from.                                                                                              |
| Source text                   | A source text is a linguistic document. This is the document that translation professionals translate from.                                                                                              |
| Target language               | A target language is a natural language. This is the language that translation professionals translate to.                                                                                                |
| Target text                   | A target text is a type of document. This is the document that translation professionals translate to.                                                                                                |
| Terminology                   | A terminology is a linguistic collection. This is the collection of words or longer sentence components that belong to a special field.                                                                     |
| Terminology extraction        | Terminology extraction is a linguistic process. This is the process of manually or automatically identifying terminology in texts.                                                                            |
| Terminology management system | A terminology management system is a computer program. This is a program that allows translation professionals to systematically store and quickly retrieve words and longer sentence components in one or more languages. 
**NOTE:** A terminology management system is particularly effective in combination with a translation-memory system. |
### Definitions, Continued

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td>Translation is a linguistic process. This is the process of transferring information from a source language to a target language in written form.</td>
</tr>
<tr>
<td>Translation memory system</td>
<td>A translation memory system is a computer program. This is a program that allows translation professionals to reuse their previous translations.</td>
</tr>
<tr>
<td></td>
<td><em>NOTE:</em> A translation memory system is particularly effective in combination with a controlled language, a terminology management system and a machine translation system.</td>
</tr>
</tbody>
</table>
Machine Translation Basics

What Is Machine Translation?

Machine translation is a linguistic process. This is the process of translating a source text into a target text where a computer program performs the actual translation.

Are All Machine Translation Systems the Same?

Today there are approximately 400 commercial machine translation systems available on the market. These 400 products can be divided into five groups of machine translation technology:

• Direct Machine Translation
• Rules-based Machine Translation
• Example-base Machine Translation
• Statistical Machine Translation
• Interlingua Machine Translation

Because of the diversity of products and technologies, it is very difficult to make general statements about machine translation.

How Does Machine Translation Work?

Rules-based machine translation systems do not simply translate word for word. Instead, rules-based machine translation systems use a sophisticated repository of grammar rules for both source language and target language. In addition, rules-based machine translation systems also use one or more dictionaries. Grammar rules and dictionaries enable rules-based machine translation systems to analyze source sentences and generate equivalent target sentences.

Fig. 1: Example of a machine translation system. The top section of the window shows the source text (German) and the bottom section shows the target text (English).

NOTE: When I use the term “machine translation” in the following text, I generally mean “rules-based machine translation”.

Continued on next page
Are Translation Memory Systems a Type of Machine Translation System?

No. A translation memory system only allows translation professionals to reuse their previous translations. A standard translation memory system cannot translate new source text. Below is an overview of how translation memories handle different translation scenarios:

<table>
<thead>
<tr>
<th>Source-text status</th>
<th>Response of translation memory system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source sentence has been previously translated (100% match)</td>
<td>Proposes previous translation.</td>
</tr>
<tr>
<td>Similar source sentence has been previously translated (partial match)</td>
<td>Proposes previous translation, highlights differences in source sentence.</td>
</tr>
<tr>
<td>New source sentence (no match)</td>
<td>No proposal.</td>
</tr>
</tbody>
</table>

Fig. 2: Example of a translation memory. The top section of the window shows the lookup component (Trados) of the translation memory. The bottom section shows the editing component (Word) of the translation memory.
## Five Machine Translation Challenges

### Many Machine Translation Implementations Have Failed

Machine translation technology promises fast and cheap translations. However, this technology is not easy to implement. Even the most advanced products available today present the user with a number of challenges. It is a fact that many well-funded and well-supported machine translation projects have failed. The primary reason for these failures is that neither vendor nor buyer were able to address the following challenges.

### Challenge 1: Dictionary Creation

One of the requirements machine translation systems have to produce high translation quality is the availability of comprehensive dictionaries. However, the following factors prevent most organizations from creating comprehensive dictionaries:

- Lack of understanding of this requirement
- Lack of expertise in the area of automatic terminology extraction
- Lack of funds to hire linguists for manual terminology extraction
- Lack of time to perform manual terminology extraction

### Challenge 2: Translation Quality

To be able to produce high translation quality, machine translation systems have a number of requirements (Example: availability of comprehensive dictionaries, controlled or pre-edited input, efficient post-editing environment). However, many vendors and buyers lack the understanding of what these requirements are and how to meet them. As a result, many machine translation systems deliver low translation quality.

### Challenge 3: Learning Capability

Some high-end machine translation systems enable users to modify the grammar rules of these systems. However, many organizations cannot afford a staff of highly skilled linguists who are dedicated to the maintenance of a machine translation system. As a result, most users of machine translation systems must correct the same error more than once. This practice is highly inefficient.

*Continued on next page*
Five Machine Translation Challenges, Continued

**Challenge 4: Rejection by Translation Professionals**

Even under ideal conditions, machine translation systems cannot fully replace human translation professionals. In a typical machine translation environment human translation professionals post-edit the output of machine translation systems. However, the following factors prevent most translation professionals from enjoying their work in a typical machine translation environment:

- Low translation quality
- Lack of an efficient post-editing environment

**Challenge 5: High Cost of Ownership**

Organizations that have been using machine translations systems successfully are paying a high price for the availability of such a system. Organizations such as Caterpillar, the European Commission, and SAP have a sizable staff dedicated to the maintenance of a machine translation system. However, such a heavy and ongoing investment in the machine translation infrastructure has a negative effect on the cost-effectiveness of a machine translation implementation.
Five Machine Translation Solutions

Machine Translation Can Be Powerful and Cost-effective

The author has successfully completed numerous machine translation projects. He has been trained by one of the leading vendors of machine translation systems (Logos) and has experience with many other machine translation products. The verdict is: Machine translation is not simple, but it does work! Machine translation can be an extremely powerful and cost-effective tool in the right environment.

Solution 1: Automatic Terminology Extraction

Once again: The availability of comprehensive dictionaries is a key requirement for any successful machine translation implementation. The author has developed one of the most powerful solutions for automatically extracting. Automatic terminology extraction has the following effects on the dictionary creation process:

- Dramatic reduction of the time required to build a comprehensive dictionary (build 100 entries per day instead of a dozen)
- Dramatic reduction of the financial resources required to build a comprehensive dictionary (higher productivity results in lower cost)

Fig. 3: Example of a terminology list that the author’s proprietary TERMinator tool created automatically.

Continued on next page
Five Machine Translation Solutions, Continued

Solution 2: Controlled Language

It is a well-known fact that machine translation systems cannot handle ambiguous, complex, and ungrammatical source text. It is equally true that most human readers of ambiguous, complex, and ungrammatical source text will also have problems following such a text. This is why the introduction of a controlled language that reduces the flaws of typical technical literature benefits primarily the customer. However, not all controlled languages are equally beneficial for translation. The author has developed CLOUT (Controlled Language Optimized for Uniform Translation), a controlled language that results in source text for easy human and machine translation.

Solution 3: Integration of Machine Translation and Translation Memory Systems

Most machine translation systems are stand-alone products. This fact has a number of negative effects on translation productivity, translation quality, and the acceptance of such systems by translation professionals.

The author has developed a unique process for integrating any machine translation system into the Trados translation memory environment. The integration of machine translation into the leading translation memory system has the following effects on the translation workflow:

• Translation professionals work in the environment they are used to.
• The translation memory provides powerful tools for post-editing machine translation (Example: spell checker, grammar checker, integrated dictionary, conversion utilities, etc.).
• The translation memory ‘remembers’ edits and eliminates the need to correct the same error twice.
• The translation memory analyzes the source text and only sends sentences to the machine translation system for which there is no human translation available. This practice eliminates any duplication of effort.
• The fact that every translation passes through the translation memory systems ensures stylistic consistency.
• Depending on the setup, the integrated system will provide translation professionals with translation proposals for all or many previously untranslated sentences. This function can have a dramatic effect on translation productivity (my tests indicate a productivity gain of 40%).

Continued on next page
Five Machine Translation Solutions, Continued

Solution 3: Integration of Machine Translation and Translation Memory Systems (continued)

Fig. 4: Example of the integration an unsupported machine translation system (Personal Translator 2002) into Trados.

Solution 4: Input Filter

Machine translation systems do not handle long sentences very well. The efficient solution is to create source text in a controlled language. However, implementing a controlled language can be a complex task. The author has developed a simple solution that reduces the percentage of unusable machine translations by simply filtering long source sentences. The fact that human translation professionals will have to deal with much less machine translation ‘junk’ will certainly improve their acceptance of machine translation.

Fig. 5: Example of an implementation of a custom-designed input filter for machine translation.

Continued on next page
The availability of comprehensive dictionaries is a key requirement for any successful machine translation implementation. However, there are two types of machine translation dictionary:

- **System dictionaries:** Every machine translation system has its own dictionary tool. These tools allow skilled terminologists to enter complex linguistic information for each word. It is true that coding these entries improves the translation quality a given machine translation system will produce. But it is also true that coding these entries is a very time-consuming and difficult task. In addition, there is no simple way of using the information stored in a system dictionary outside a given system. In other words, it is not possible to export this information to another machine translation system or to a terminology database for human translation professionals.

- **Import dictionaries:** Most machine translation systems have a dictionary import tool. These tools allow translation professionals to enter simple linguistic information for each word. It is true that these entries improve the translation quality of a given machine translation system not as much as entries in a system dictionary. But it is also true that translation professionals can create these import dictionaries from existing terminology databases. In other words: It is possible to create import dictionaries at the push of a button. It is also possible to use these import dictionaries with more than one machine translation system.

Fig. 6: Example of a multi-window system dictionary (left) and a simple import dictionary (right).
Five Machine Translation Solutions, Continued

Solution 5: Selective Dictionary/Rules Work (continued)

Some high-end machine translation systems allow the modification of the rules repository to adjust the machine translation system to a specific writing style. It is true that coding these rules improves the translation quality a given machine translation system will produce. But it is also true that coding these rules is a very time-consuming and difficult task. In addition, there is no simple way of using the information stored in a rules repository outside a given system. In other words, it is not possible to export this information to another machine translation system.

It is for these reasons that the author does not recommend following the standard practice of coding rules. Instead, the author recommends either using a controlled language or an input filter. Both of these solutions are vastly more cost-effective than doing product-specific coding work.
Implementation Scenario 1

Overview

This scenario is a description of the most efficient translation workflow available today. In this scenario the authoring stage is fully adjusted to meeting the needs of the translation stage. However, a streamlined authoring-translation process requires a substantial investment in tools, training, and workflow design.

Tools

The following areas will have requirements for new tools:
• Terminology Management

<table>
<thead>
<tr>
<th>Type</th>
<th>Commercial Products</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminology Workflow</td>
<td>MultiTerm Server</td>
<td>Implemented</td>
</tr>
</tbody>
</table>

• Authoring

<table>
<thead>
<tr>
<th>Type</th>
<th>Commercial Products</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled Language Style</td>
<td>acrolinx</td>
<td>approx. $100.000*</td>
</tr>
<tr>
<td>Checker</td>
<td>multilint (support for</td>
<td>no quote yet</td>
</tr>
<tr>
<td></td>
<td>Arbortext not clear)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MaxIt</td>
<td>approx. $1.000 per seat*</td>
</tr>
<tr>
<td></td>
<td>(currently supports</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arbortext)</td>
<td></td>
</tr>
</tbody>
</table>

This quote is for the Microsoft Word client. A client for Arbortext is under development and is scheduled for release in two months. acrolinx states that they can also port their client to a software development platform.

• Translation

<table>
<thead>
<tr>
<th>Type</th>
<th>Commercial Products</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-end Machine</td>
<td>comprendid</td>
<td>no quote yet</td>
</tr>
<tr>
<td>Translation Suite for</td>
<td>IBM WebSpere</td>
<td>approx. $ 80.000</td>
</tr>
<tr>
<td>Multiple Languages</td>
<td>Translation Server</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Logos</td>
<td>no quote yet</td>
</tr>
</tbody>
</table>

Continued on next page
Implementation Scenario 1, Continued

Workflow
- Authors (documentation and software) use style checkers to create source text.
- All terminology issues are resolved prior to translation.
- The translation memory system analyses the source text.
  - 100% matches: Re-use of existing translation
  - 75% matches: To human translation professional for updating
  - <75% matches: To machine translation system for translation and subsequent post-editing by human translation professional

Benefits
- Authors who use a controlled language in combination with style checkers produce text that is easy to understand. This fact will immediately improve the usability of the software products.
- As controlled source text is easy to understand, it is also easy to translate. This fact will dramatically improve the translation quality of machine translation systems.
- As controlled source text is more standardized, it is more likely that authors will re-use similar sentence constructions. This fact will improve the translation productivity of the human translation professionals as they will find more matches than with uncontrolled source text.
- Using controlled source text in combination with machine translation will have a dramatic effect on the business aspect of translation:
  - Translation costs will go down
  - Productivity will go up
  - Translation quality/consistency will also go up

Drawbacks
- This is a complex solution that will involve a major investment in tools and training.
- It will take months if not a year to implement this scenario.
- It will take an effort to overcome resistance on the part of authors and translators.
Implementation Scenario 2

Overview
This scenario is a description of an efficient translation workflow the implementation of which will require a relatively minor effort. In this scenario the authoring stage is left as is. However, the introduction of machine translation will change the translation workflow.

Tools
The following areas will have requirements for new tools:

- Terminology Management

<table>
<thead>
<tr>
<th>Type</th>
<th>Commercial Products</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminology Extraction</td>
<td>muegge.cc</td>
<td>Minimal</td>
</tr>
<tr>
<td></td>
<td>TERMINator*</td>
<td>Implemented</td>
</tr>
<tr>
<td>Terminology Workflow</td>
<td>MultiTerm Server</td>
<td></td>
</tr>
</tbody>
</table>

*This tool is currently not packaged as a product (no documentation, etc.)

- Translation

<table>
<thead>
<tr>
<th>Type</th>
<th>Commercial Products</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Rules-based Machine Translation Systems</td>
<td>Personal Translator for German and French prompt for Russian many others available</td>
<td>$300 per language</td>
</tr>
<tr>
<td></td>
<td>LanguageWeaver xplanation</td>
<td>$60.000</td>
</tr>
</tbody>
</table>

Continued on next page
Implementation Scenario 2, Continued

**Workflow**

- Translation professionals extract terminology prior to the translation process to build a comprehensive dictionary for human and machine translation.
- For rules-based machine translation: Translation professionals either pre-edit the source text for machine translation or use an input filter to avoid mistranslations by the machine translation system. For example-based machine translation: We are currently in the process of testing one system. At this point, we assume that there will be no preparatory work. However, we will need to validate this assumption.
- The translation memory system analyses the source text.
  - 100% matches: Re-use of existing translation
  - 75% matches: To human translation professional for updating
  - <75% matches: To machine translation system for translation and subsequent post-editing by human translation professional

**Benefits**

- As this scenario focuses on the translation stage, implementation can be faster and easier than in the other scenario.
- The implementation costs will also be much lower, especially when working with individual rules-based machine translation systems.
- Using machine translation will have a significant effect on the business aspect of translation:
  - Translation costs will go down
  - Productivity will go up
  - Translation quality/consistency will also go up

**Drawbacks**

- The benefits of this solution will be limited to the translation stage. There will be no positive effect for customers who use the English documentation.
- This scenario is not nearly as efficient as the other scenario. Human translators will spend more time for post-editing.
- It will take an effort to overcome resistance on the part of translators.
About the Author

Industry Experience

Ten years of service in the language industry in a wide variety of roles ranging from technical author to translator and terminologist to tester and trainer.

- Acted as language technology consultant for AmdocsClarify, DaimlerChrysler, and SAP.
- Taught Translation Management & Terminology at the postgraduate level.
- Developed a number of unique translation solutions. The German Patent office has accepted a patent application for my terminology extraction solution.

Education

Currently Ph.D. program in Applied Linguistics
University of Leipzig, Germany

1996-1998 MA program in Translation
Monterey Institute of International Studies

1986-1991 MA program in Telecommunication
University of Oregon

1983-1986 BA program in English Linguistics and Literature
University of Tuebingen, Germany

Selected Publications


