[*Translating and the Computer 20*. Proceedings from Aslib conference, 12 & 13 November 1998]

Improving Translation at the Source

Dawn Murphy, Jane Mason, Stuart Sklair
Multilingual Technology Ltd (MTL)
Eaton House
Wigmore Lane
Luton, Beds
LU2 9EZ
Tel: +44 (0)1582 702000
dmurphy@multilingualtechnology.com

Cost, quality and time, not necessarily in that order, are the three most important factors in a company's translation strategy. Or, more pertinently, how to *cut* costs, how to *improve* quality and how to *save* time.

Improvements in the quality and efficiency of translation can be effected at the very birth of a document, that is, at the time of authoring. There are several ways of ensuring consistency and quality of authoring to various degrees, all of which have a beneficial effect at the translation stage.

This paper examines the concepts of controlled authoring and author memory and the supporting technologies, the processes involved in developing such techniques within an organisation and the business benefits that such techniques bring.

Introduction

Controlled languages and translation technology

There has been a great deal of interest recently in controlled languages, especially within the automotive and telecommunications industries. The attendance at the CLAW conference in Pittsburgh this year and in Leuven two years ago reveal the extent of this growing interest. At the same time machine translation is becoming accepted in some industries as a worthwhile investment, as experiences at Océ[1] and Mitel[2] show. The other major translation technology, translation memory, has been in use for several years now and it is widely accepted in business that great savings can be made from the leverage of previously translated material. One company, Nortel, has been influential in setting up a cross-industry National Consortium which aims to advance the development of both controlled language and computer-assisted translation tools, define standards and evaluation criteria, and promote the use of such tools.

Companies have accepted that translation memory and terminology database tools can be worth investing in for large volumes of revised text such as updates to workshop

[1] Implementing MT at Oce. Language International 9.6 (1997), pp 16-17

[2] Machine Translation Finds a Home at Mitel. Language International 10.3 (1998), pp 40-41

manuals, although there is still a great reliance on traditional human translators. People are generally more sceptical about machine translation, due to the proliferation of stories in the press about ridiculous results obtained from MT systems, usually because they were given the wrong sort of text in an uncontrolled fashion.

Why use machine translation?

Using fully or semi- automated translation is more important than ever now that companies are expanding into new emerging markets such as Russia, China and Latin America, and are therefore having to translate into new languages or language variants, such as Brazilian Portuguese and Mexican Spanish. The rate of technological advances, particularly in the IT industry at least, is creating products with extremely limited shelf life. The result of this may be that new products are launched within months of the previous version or model. This of course has an impact on translation, as the documentation accompanying such products has to be ready at the time of product launch and should also ideally be ready in every target language.

According to Rose Lockwood in her article "Global English and Language Market Trends"[3], English is likely to "continue to be either a lingua franca, or else a pivot language for international communication." However, in an increasingly competitive global market, it is more important than ever for businesses to translate into the language of their customers. It is imperative therefore that businesses use the technologies available to their full potential, in order to keep translation costs down while translation output increases. Lockwood also points out that the effect of the continuing status of English as a commercial language is that more languages will be paired with English in translation tools.

Concepts

Controlled authoring

Controlled authoring is the application of terminology, style and grammar rules, above and beyond the normal rules of the language, in order to simplify and standardise the structure and words used. These rules reduce ambiguity, and increase consistency of expression and terminology. Such rules might govern the number of nouns allowed in a compound noun phrase, the maximum length of sentences, or the tense of verbs in instructions.

This control may be carried out for several reasons. Firstly, to make the source text clearer and therefore easier for non-native speakers to understand. Secondly, to make the source text clearer for translators to understand. And thirdly, to make the source text easier to process by machine.

The first form of controlled language, AECMA Simplified English (AECMA SE)[4], was developed for the aerospace industries. AECMA SE was never intended to be translated, but was intended to make documentation less open to mis-interpretation by

[3] Lockwood, R**.** Global English and Language Market Trends**.** Language International 10.4 (1998), pp 16-18

[4] AECMA (1996) A Guide for the Preparation of Aircraft Maintenance Documentation in the Aerospace Maintenance Language. AECMA Simplified English**.** AECMA Document: PSC-85-16598, Issue 1. Brussels.

non-native speakers of English. Caterpillar[5], Boeing[6] and General Motors[7] on the other hand, have all invested in controlled language development with the aim of obtaining better results from machine translation or translation memory tools.

Author memory

Author memory, compared with controlled language, is a relatively new concept. The idea came from translation memory. Previously-authored material is held in the form of phrases or segments of text, which pop up as the author writes and can be reused without the need to retype the whole segment. To ensure that the hits found in author memory and presented to the author are not too numerous or irrelevant, controlled language can also be used to maintain precise, consistent text. The more precise the text is, the more likely it is that the most relevant match is found in memory.

Language Variant Management

Many companies, especially in the automotive and aerospace industries, are currently authoring in SGML or intend to do so in the future. It is possible to make use of this method of authoring to obtain the maximum possible re-use of previously authored and translated material. Language Variant Management is a way of storing SGML components, such as operations, procedures, and headings, and linking these components to their translations in a document management system. Translation memory can then be used to identify similar sentences at a finer level of granularity to the SGML component. The effectiveness of Language Variant Management can be increased by using it in combination with controlled authoring to ensure consistent text and consequently a high hit rate for matching SGML elements.

Processes

Authoring tools

The only way to effectively implement a controlled language or controlled vocabulary is to use checking tools. If the author is simply provided with a set of rules to follow or a list of terminology to use, there is too much room for human error. An errant hyphen within a term can mean a sentence is not retrieved from translation memory for instance. Controlled terminology checkers extract terms from a text, check them against a database for approved, unapproved or unknown status, and highlight to the author where an unapproved or unknown term has been used. The tool may also suggest an approved alternative in the case of an unapproved term being used. In the case of an unknown term, the author may make a formal request for that term to be added to the controlled vocabulary.

[5] Hayes, P., Maxwell, S. & Schmandt, L. (1996) 'Controlled English advantages for translated and original English documents' in Proceedings of the First International Workshop on Controlled Language Applications (CLAW 96). 1996, Leuven, pp 84-92.

[6] Wojcik, R.H. & Holmback, H. (1996) 'Getting a controlled language off the ground at Boeing' in Proceedings of the First International Workshop on Controlled Language Applications (CLAW 96). 1996, Leuven, pp 22-31.

[7] Means, L. & Godden, K. (1996) 'The controlled automotive service language (CASL) project' in Proceedings of the First International Workshop on Controlled Language Applications (CLAW 96). 1996, Leuven, pp 106-113.

Controlled language checkers go one step further and check both vocabulary and syntax or style against the controlled language rules. They may inform the author which rule has been violated, and suggest an alternative replacement. Both controlled language checkers and controlled terminology checkers can be used interactively as the author writes as a form of reference tool or self-editing tool or in batch mode as a form of post-editing.

Author memory tools plug in to the authoring environment, and prompt the author as a phrase is written, offering possible phrases from memory. Author memory can be used in conjunction with a terminology database lookup tool to create an "authors' workbench".

Developing controlled authoring

Developing a controlled vocabulary

There are essentially five stages in the development of controlled authoring within an organisation. Firstly, the terminology and vocabulary used by the authors and translators must be tackled. This would involve analysing the material currently being authored within the company, extracting the terminology used in this material, and identifying potential issues with the terms. Those issues might be:

- one word having two different meanings e.g. *replace,* which could mean either to remove something and then put it back, or to remove something and put something else in its place.
- two different terms for one concept e.g. *windscreen* and *windshield*
- inconsistent orthography of terms e.g. *locknut* and *lock nut*
- complex structure of terms e.g. the compound *outlet hose to the thermostat,* which within such sentences as *Re-connect the outlet hose to the thermostat,* would cause structural ambiguity. It would not be clear either to a human translator or an MT system whether the sentence is saying that the *outlet hose* should be re-connected to the *thermostat,* or that the *outlet hose to the thermostat* should be re-connected to something else.

Eventually, the list of terms would be reduced to a definitive set of unambiguous terms with one term per concept.

Developing a controlled language

The next step is to develop the controlled language rules. Again, the documents currently authored are analysed for potential sources of ambiguity and for complex structures which might prove difficult for a MT system to process. A set of rules defining for instance the maximum length of sentences, maximum number of clauses in one sentence, and so on, is devised.

Selecting an authoring tool

Following the development of the controlled terminology and language rules, the logical step would be to investigate technology to support the implementation of the rules. However, this stage could equally be carried out after the controlled terminology has been developed and before moving on to developing fully controlled language.

The selection of an appropriate authoring tool involves identifying the requirements for the company's specific authoring environment, evaluating any tools on the market, and adapting any selected tool to the company's requirements. Most authoring tools are sold as bespoke tools, and therefore will be customised by the developer for the buyer. This allows a certain degree of flexibility - for instance, a tool which is only currently available in Microsoft Word can still be a viable option to a company which authors in Interleaf.

Implementation

The next stage is the implementation of the controlled language. This is often the hardest part. Authors must be convinced of the benefits of using the controlled language, as they will find it extremely intrusive at first. A certain amount of training is required and a pilot is needed, both to verify that the controlled language is workable and that the vocabulary is comprehensive enough, and also to check out the suitability of the chosen controlled language checking tool.

Maintenance

Finally, continuous management of updates to the controlled vocabulary and of requests for new terminology will be needed.

Results

Controlled authoring used with TM

Controlled language can improve the hit rate for translation memory because firstly, sentences are more likely to be written the same way due to the restricted syntax. Secondly, the same terminology will be used due to the restricted terminology rules. For instance, Figure 1 shows a paragraph taken from a car workshop manual.

> *With transaxle at normal operating temperature, park vehicle on level surface. Place gear selector in P or N position and apply handbrake. Allow engine to idle. Remove dipstick and check that fluid level is between its marks.*
>
> **Figure 1**

In Figure 2, we show how the same paragraph could have been written a different way, say for a different model by a different author. We can see that if this piece of text was put through a translation memory system, the hit rate for exactly matching text would be nil.

> *With the transaxle at normal temperature, park the vehicle on a level surface. Place the gear selector in P or N position and apply the parking brake. Allow the engine to idle. Remove the dipstick. Check that the transmission fluid level is between the marks on the dipstick.*
>
> **Figure 2**

If author memory is used, or controlled language and terminology rules are applied, it is much more likely that the two authors will produce similar text, or that an author will write similar text in two different manuals. For instance, the terminology rules which could be applied here are:

1. Unacceptable term *normal temperature* - > Acceptable term *normal operating temperature* (be specific about what sort of *temperature.)*

2. Unacceptable term *fluid level* - > Acceptable term *transmission fluid level* (there could also be other types *of fluid level* such as *coolant level).*
3. Unacceptable term *handbrake* - > Acceptable term *parking brake* (this could equally be the other way around, with *handbrake* being the acceptable term.)

This would ensure consistency of terminology across different products and different authors. To ensure the best possible hit rate from translation memory though, author memory can be used to prompt the author to use previously written text. Alternatively, controlled language rules such as the following could be applied:

1. Only use approved terminology.
2. Do not convey more than one instruction per sentence.
3. Do not omit articles.
4. Do not use *it* or *its* - specify the noun.

If both authors follow these rules, the following text should be produced in both cases:

> *With the transaxle at normal operating temperature, park the vehicle on a level surface. Place the gear selector in the P or N position. Apply the parking brake. Allow the engine to idle. Remove the dipstick. Check that the transmission fluid level is between the marks on the dipstick.*

Figure 3

Controlled authoring used with MT

Controlled language can vastly improve the quality of MT output as it reduces or eliminates part of speech ambiguity, semantic and structural ambiguity, anaphora and ellipsis. The following piece of text has instances of missing articles e.g. *Remove sump,* missing nouns e.g. *and dry,* use of pronouns i.e. *front ones,* and use of phrasal verbs i.e. *pour out.*

> *Remove rear sump bolts and slacken front ones. Carefully lower sump and drain as much fluid as possible. Remove sump and pour out remaining fluid. Remove filter. Clean oil sump in solvent and dry.*

Figure 4

We used a leading MT system, with a customised dictionary, to translate the text into French, and the resulting translation was this:

> *Enlever les boulons arrière de carter d'huile et dégager I'avant ceux. Abaisser soigneusement le carter de vidange et vidanger autant liquide que possible. Enlever le carter de vidange et verser hors du liquide restant. Enlever le filtre. Carter de vidange de pétrole raffiné dans dissolvant et sec.*

Figure 5

There are several places where the MT system has misinterpreted the English and consequently produced a faulty or meaningless French translation. For instance, when translating the sentence *Clean oil sump in solvent and dry,* the missing article before *oil sump* means that the system has less information to help it disambiguate the word *clean,* which could be a verb or an adjective. Similarly, the word *dry* has been mistranslated as an adjective instead of a verb, because the noun which would normally follow the verb has been omitted.

Using a few typical controlled language rules, we rewrote the text, and translated it again using the MT system. These were the controlled language rules we used:

1. Do not omit articles, such as *the* and *a.*
2. Do not use pronouns instead of nouns.
3. Do not try to convey more than one instruction per sentence.
4. Do not use phrasal verbs, such as *pour out.*
5. Do not omit implicit nouns.

Figure 6 shows the rewritten text and Figure 7 the result – a much improved French translation, which now only needs some very minor post-editing.

> *Remove the rear sump bolts. Slacken the front sump bolts. Carefully lower the sump. Drain as much fluid as possible. Remove the sump. Discard the remaining fluid. Remove the filter. Clean the oil sump in solvent. Dry the oil sump.*
>
> **Figure 6**
>
> *Enlever les boulons arrière de carter d'huile. Dégager les boulons avant de carter d'huile. Abaisser soigneusement le carter de vidange. Vidanger autant liquide que possible. Enlever le carter de vidange. Jeter le liquide restant. Enlever le filtre. Nettoyer le carter d'huile dans le dissolvant. Sécher le carter d'huile.*
>
> **Figure 7**

Business Benefits

There are many business benefits to be gained from the implementation of controlled authoring and authoring technology.

Firstly, the cost of creating multilingual documentation can be considerably reduced with the use of translation memory. Using authoring technology as well can reduce these costs even further by increasing the level of reuse. Author memory also reduces authoring time and the associated costs. It is difficult to estimate the potential savings without considering the precise nature of a company's documentation, the level of re-use, the number of languages the documentation is translated into, the number of pages produced each year and the type of authoring or translation technology already used. However, recent studies by MTL in the automotive sector have suggested that introducing a combination of language variant management, author memory and controlled authoring could reduce translation costs by as much as 20-30% in addition to savings already made from the use of translation memory.

Without controlled authoring, machine translation for publication purposes often requires too much post-editing, and is therefore usually not a viable option. With controlled authoring, machine translation produces fewer errors in the raw output, and therefore translation costs can be reduced even further. Documentation that previously could only be produced in the source language because of the costs involved, may now be translated. Additional target languages may also be possible.

The turnaround time for creating multilingual documentation can also be greatly reduced with the use of authoring tools. Using controlled authoring can improve the clarity and comprehensibility of the text for the translator, and if controlled terminology lists are provided with target language equivalents, the number of translators' queries during translation will be reduced. One of the main delays in producing a translation is the constant "to-ing and fro-ing" between the translator, the project manager and the document creator to resolve queries over textual ambiguities and unknown terminology. Controlled authoring, author memory and language variant management also make more re-use possible and so reduce the amount of text which has to be translated.

In cases where a market is too small to justify translation costs, improved quality of source documentation can go some way to meeting the customer's need. The documentation may still be in English, but at least it will be in clear, easy to read, unambiguous English. In addition, in the case of, for example, a service manual being produced in English for non-English speakers, a bilingual glossary and terminology browsing tool could also be shipped with the manual as a comprehension aid to help the reader without the need for full translation.

These days, both authoring and translation are often outsourced to agencies, often to several different agencies and in different countries. This means it is harder than ever to maintain consistency of terminology and style across different documents. Implementing a combined programme of controlled authoring, authoring technology and translation technology allows a global company to maintain its corporate image in its documentation.

Finally, if the quality of source documentation is improved, there is less room for mis-interpretation of instructions by end-users or by translators. This can mean less maintenance and repair costs. Furthermore, there is a reduced risk of litigation from ambiguous, unclear and possibly even inaccurate documentation.

Conclusion

We will need to get past certain barriers of artificial intelligence before MT will be as good as the human translator. This will happen one day, but that day is a long time in the future. We need therefore to harness the technology now by working with it. Controlled authoring is a way of doing this. But controlled authoring also has the useful side effects of improving the source document for readers in the source language, plus making translation easier, cutting down on translators' terminology queries and allowing them to concentrate on the language they are translating into rather than deciphering the language they are translating out of.