In *Papers from the XIII VAAKKI symposium 1993*, Vaasa.

WHEN AND WHY SHOULD TRANSLATIONS BE REUSED?

Magnus Merkel
Department of Computer and Information Science, Linköping University

Med fler och fler verktyg för översättning tillgängliga kommer antagligen översättare att bli betydligt mer effektiva än idag. Men det är inte enbart själva översättningsprogrammen som är av intresse. Effektiva analyser av källtexten och av ev. tidigare måltexter har också avgörande betydelse, särskilt när det gäller att välja mellan en mängd olika typer av verktyg. Vissa texter lämpar sig väl för översättningsminnesbaserad översättning (translation memory) medan andra texter kräver manuell översättning rakt igenom. Om översättningen sker inom ett område där man har tillgång till gamla översättningar bör man kunna analysera den gamla och nya källtexten och få ett mått på om det är värt att återanvända den tidigare översättningen. Ett exempel på återanvändning är att använda s k alignment-program, som parar ihop meningar i käll- och måltexten. I rapporten finns exempel på resultat från automatiska textanalyser av teknisk dokumentation och exempel på hur man kan skapa översättningsminnen av redan översatt material.

Introduction

In this paper I would like to discuss the situation at hand when a translator, or team of translators, is faced with a relatively long document that is going to be translated into another language as quickly and as accurately as possible. With modern technology available, this situation could be very challenging. If the translator(s) had a versatile toolbox filled with translation software of different kinds, it would seem like the problem would be simple. The answer would be to pick the appropriate tool out of the toolbox and thereby make the translation work smooth and efficient. However, given that this translation toolbox existed, it would not solve all the problems, instead it would introduce the task of deciding what tool to use for a particular document.

It is claimed here that a translator must be able to evaluate a given source document before the actual translation work starts. In particular the translator must be able to decide whether translation tools can and should be used when a given document is translated or whether it should be left to manual translation.

A related question is whether old translations should be reused in, for example, new versions of handbooks. To answer such a question, it would be necessary to have reached some measurements of how alike the old and the new versions of the documents are, and, if this likeness could be utilised by a translation tool. If translations could be reused on a large scale basis, it would have implications on the cost, speed and quality of the translations. I think it is relatively safe to say that reusing translations would decrease costs and increase speed of the translation, but the quality aspect is more problematic.

What is quality in translation work? There is probably a core of characteristics of good translations that would apply for most kinds of text, such as correct translation of content, correct spelling, correct terminology, etc. But there are other aspects, such as variation in style and phrasing, or consistency on different levels, that have different applications in different types of text.

In the paper I first discuss the characteristics of technical documentation and then I go on to the implications of different types of translation methods. Then there is a section on tools that would help translators in their work. The first is a text analysis tool, or perhaps better named, a recurrency analyser. The second is an alignment tool that analyses a source and target text and constructs a database of translated sentences, which can be reused.

Technical documentation

A large proportion of commercial translation is done within the field of technical documentation, that is manuals, instructions, descriptions of machinery, etc. The technical documentation attached to a technical product is getting more and more important in many aspects. First of all, and perhaps most significant, a product cannot be shipped without technical documentation. For many products and markets, this means that the documentation has to be translated. Delays can cause serious losses of market shares within business fields where the technology development is fast. Time is therefore a critical aspect for companies in need of technical translations.

From the translator’s point of view, technical documentation is getting bulkier and bulkier. As noted above, companies ask for higher translation output, because of the harsh competition on the market. To be able to cope with the demands from industry, translators must work together in teams. The days are gone when one translator could translate a whole set of technical handbooks belonging to one particular product. It is not unreasonable to think that eight or more people could be working on the same documentation. But to be successful, a team of translators needs co-ordination and organisation. The translation co-ordinator has a formidable task of getting an accurate overview of the actual translation work at any given point.

In my view, the goal to translate quickly, cheaply and with good quality, is best achieved if the co-ordinator has detailed information about the text to be translated. The co-ordinator’s task is then to distribute this knowledge to the translators in the team and also find a way of keeping everybody up-to-date with what problems are being solved, etc. This is hard to achieve by traditional working methods, but should be possible if new and existing tools were developed and taken into use.

Translation methods

Before presenting the tools, I would like to develop the discussion on the types of translation methods involved in the translation of technical documentation. As I see it there are three major translation alternatives at disposal today: 1. Manual translation, 2. Rule-based machine translation and 3. Memory-based translation.

The manual translation situation holds the problem of consistency of terminology, phraseology and style when a team of translators are working on a large documents. See also the previous section.

By rule-based MT systems I mean systems that have large multi-lingual grammars and lexicons and that run in batch to be post-edited at a later stage. Apart from being large and complex, these systems are expensive. They are also difficult to develop and maintain as it is difficult to foresee how certain changes in the linguistic knowledge bases would affect the behaviour of the system. However, for certain text types, rule-based MT have been claimed to be very cost-efficient (see, for example, Slocum, 1987).

The memory-based translation tools are to be considered as machine-aided translation where the translator is still in control. The idea is that translations of a text segment should be reused when the segment reappears in the document. Translated segments are stored in a database and can be retrieved by the system when needed. This type of system is relatively inexpensive and does not require a phase where large knowledge bases have to be created by coding grammars or lexicon. Instead the translation memories are built up as you translate, and the larger your database is for a certain text type and domain, the larger the gain is. Memory-based translation directs translators by default towards consistency. Commercial systems that can be mentioned here are IBM's Translation Manager/2 and Trados TWB. Both these systems are based on the same idea.

Recurrency

One important aspect of making technical translations more efficient, is how repetitious, or recurrent, a document is *internally*, that is, if whole segments like paragraphs, sentences and phrases recur within the same document. Another aspect is whether there is recurrency across two or several similar documents, for instance, a new version of a handbook in relation to a previous version. The latter type is here called *external* recurrency.

The most superficial level of recurrency is the repetitions of exact strings. This requires no advanced NLP techniques, it is just a matter of string processing. The next step would be to introduce some kind of fuzzy matching of recurrency. Here wild cards or variables are used to represent parts of strings that do not have to be identical, but could nevertheless be interesting. To illustrate these two kinds of matching, consider the following two examples:

1. "Choose the OK button."
2. "Choose the Exit button."

If we adopted the first approach of exact matching, string 1 above would only be matched by exactly the same strings of characters as are given inside the quotes. However, if the fuzzy matching technique would be used, the second string would be matched by the first string, as well as strings like "Choose the Cancel button.", "Choose the Help button.", etc.

We could go higher up on the abstraction ladder and use formal grammars to represent the text. If we did we would be able to find patterns of syntactic constructions, grammatical features such as co-ordination, sentence mode, tense changes, etc. The more abstract our representation would be the more patterns we would probably find. However, the higher abstraction level would definitely "cost more" in terms of computational power and sophistication of the grammatical knowledge bases.

Ideally all levels of recurrency analysis should be accessible to a translator. But, as this is not possible at the moment, it might be interesting to consider what the simplest levels of recurrency analysis would yield.

Tool 1: The Recurrency Analyser

At Linköping University we have developed a Recurrency Analyser to measure internal and external recurrency on sentence and phrase level. A corpus consisting of 1 million words of American computer handbook texts were analysed with this tool.

The tool performs an exhaustive search of the text and the user does not have to specify what to look for. In this way it is similar to the program developed for literary analysis of repeated structures in Kingston, Canada (Lessard & Hamm, 1991). One major difference between our program and Lessard & Hamm's is that our tool does a calculation on how much of the total text that is recurring.

On the sentence level, we have analysed the whole corpus in one batch, but for phrase/string level analysis the largest text analysed in one run was around 250,000 running words. The system runs on a Sun Sparc station and is written in C.

The results show that for most of the texts in the corpus, there is a high degree of recurrency (both on sentence and phrase level). It should be pointed out that "phrases" here are the manually revised list of maximal strings that are the output from the system. The manual revision was done by myself by removing all non-phrases and phrases that I regarded did not have corresponding Swedish phrases as translations.

Here is an example of how the topmost elements of a sentence list may look like.

```
choose the ok button. 211
result    152
example   147
- or -    82
— or —    82
you can display a new result by pressing the
update field key (f9) or by choosing print
merge from the file menu.  30
do this   30
example:  26
do one of the following:   21
```

Below the highest recurrency scores in the corpus analysis are shown:

Sentence level
Internally: Up to 25 per cent of the text is made up by recurrent sentences
Externally: Up to 20 per cent of the text is made up by sentences that occur in a different document.
Phrase level (maximal strings manually revised for translation purposes)
Internally: 31 per cent
Externally: 15 per cent

Combinations
Internal sentences and phrases
43 per cent
External sentences and phrases
31 per cent
Internal/external sentences and external phrases
55 per cent

See Merkel, 1992, for a more detailed account of the Text Analysis Tool and the corpus analysis.

The implications of the recurrency analysis above are that with translation memories available, 31 per cent of one sample text would practically have been translated, given a translation tool that utilise the old translation memory. The translation would have been consistent throughout the document, no matter how many translators that had been working on it, as each translator would have been able to benefit from already translated segments. Note that the recurrency figures would have been "higher" if some method of "fuzzy" matching had been adopted. The most important implication is to regard the recurrency figures as a diagnosis of the text, namely the recurrency characteristics. If these figures are very low then it would probably be safe to apply manual translation to the text. It would certainly be out of the question to use a memory-based translation tool.

Tool 2: The Alignment Tool (LinAlign)

Another tool developed at Linköping is an Alignment program which creates translation memories of a source and target text, that is, it links a sentence in the original with a corresponding sentence in the target document. There are different techniques to the alignment of segments. Most notable has been the statistical approach, which the LinAlign tool also adheres to. The most well-known statistical algorithm is the one done by Gale & Church (1991). Ours is a modified version of their algorithm.

The algorithm is based on three assumptions of the source and target texts.

1. The source and target texts are similarly ordered.
2. If two sentences in one text are combined to one sentence in the other text, it is always adjacent sentences that have been joined.
3. The alignment is based on paragraph and sentence lengths (number of characters.

Apart from 1-1 relations, LinAlign also handles 1-2 and 2-1 relations (1 source sentence - 2 target sentences, 2 source sentences - 1 target sentence).

Below is a sample of the output from the LinAlign program.

```
f1:21.1 l  Specify the amount of time before
  you recieve messages about printer
  problems.
f2:21.1 l  Ange efter hur lång tid ett
  meddelande rörande skrivarproblem ska
  visas.

f1:22.1 l  Select the default printer.
f2:22.1 l  Välj standardskrivaren.

f1:23.1 The following sections explain how
  to perform each of these tasks.
f2:23.1 Följande avsnitt förklarar hur du
  vidtar dessa åtgärder.
```

The code before each segment gives information about each document and its respective paragraph and sentence ordering.

Alignment test

A test of the LinAlign tool when run on a manually translated text, showed that out of 624 sentences, it failed on only 4 sentences. The test was done on a English-Swedish corpus. Church & Gale (1991) reported that their tool when tested on a similarly sized English-French material failed on 22 sentences out of 621. It is of course impossible to draw any conclusions on the quality of the tools from such small and different test materials.

However, one interesting factor found when we analysed the source text with the Recurrency Analyser was that 23 sentence types were repeated between 2 to 19 times in the text. A recurrency test on the target text revealed that out of these 23 sentence types 20 had been translated with consistent translations. The 3 sentence types (all with the frequency 2) that had different translations could have had consistent translations.

Recurrent source sentences with different translations

1. The options available in the dialog box below may vary, depending on the network you are using.
1a. Vilka alternativ som finns i dialogrutan nedan beror på vilket nätverk du använder.
1b. Tillgängliga alternativ i dialogrutan beror på vilket nätverk du använder.
2. Select the port you want to assign the printer to.
2a. Markera den port du har anslutit skrivaren till.
2b. Välj vilken port du vill ansluta skrivaren till.
3. Select the port you want to use.
3a. Välj den port du vill använda.
3b. Markera den port du vill använda.

In other words, there was nothing special in the context that demanded variation. It was just what the translator had chosen at a certain point in the translation process, unaware of the fact that the exact sentence occurred at some other text segment.

Conclusions

Translators should have better tools at their disposal. However, it is not only translation software that is needed, there should also be text analysis tools that supply the translator with information on what kind of tool should be applied to a given text. All texts would not benefit from the use of translation software, but some would. The analysis tools could also compare two sets of documents and find out how similar they are, thereby giving the translator the incentive to reuse already existing translations.

Reusing old translations requires the existence of translation memories and memory-based translation software. Translation memories can be built up as-you-translate if you use the appropriate software, but it is also possible, by means of alignment program, to "recover" old translations and recycle them.

It remains to be seen what kind of effects translation memories will have on the language quality of the translations. It will mean better consistency, but will it have any effect on text coherence and text binding? The results of this study indicate no negative effects when it comes to the translation of technical documentation. However, larger multi-lingual corpora in different domains must be investigated before we can answer this question satisfactorily.

References

Gale W. A. & Church K. W: (1991), "A program for aligning sentences in bilingual corpora", in *Proceedings from ACL-91*, pp 177-184, Berkeley.

Lessard G & Hamm J-J (1991). Computer-aided Analysis of Repeated Structures: the Case of Stendhal’s Armance, in Journal of Literary and Linguistic Computing, Vol. 6, No. 4, 1991.

Merkel, Magnus (1992). Recurrent Patterns in Technical Documentation. Research Report, Dept. of Computer and Information Science, Linköping University.

Slocum J (1987). A Survey of Machine Translation: its History, Current Status, and Future Prospects. In *Machine Translation Systems*, ed. Slocum, Cambridge University Press.