>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 07: L'aplicació dels corpus lingüístics a la traducció : desembre 2009 :
http://www.fti.uab.cat/tradumatica/revista : ISSN: 1578-7559

Future Trends in Translation Memory

Daniel Benito
ATRIL

Abstract: This article looks at some of the latest advances in translation memory technology and how a corpus-linguistic approach could be applied to further extend them in order to make them more appealing. It also explores how the nature of the translation industry can affect whether new technologies are widely adopted or not.

Keywords: translation memory, bitexts, subsegments, consistency, automation, costs

Introduction

Where is TM (Translation Memory) technology headed? Since the appearance of commercial implementations in the early 1990s, the technology itself has not advanced much. The translation tools built on top of TM technology have indeed evolved considerably in the last two decades, but the improvements have been mostly concentrated in the complementary features offered by integrated translation environments, rather than significant increases in the level of reuse of previous translations. We will explore the reasons behind this stagnation, whether TM technology can move forward and, if so, where it will lead us.

To do so, let us take a quick look at the history of TM and how it evolved into its current state.

A Brief History of TM

The basic ideas behind TM technology arose in the late 1960s and early 1970s, as part of work done on “translator workstations” (Hutchins, 1998). As opposed to the utopian goal of fully automated machine translation, the aim of these workstations was to provide translators with a number of different resources that would allow them to carry out their work more accurately and efficiently. In addition to the more obvious utility of term banks and glossaries, early researchers had already envisioned the use of translation archives as a reference tool. Translators would be able to quickly consult past translations to see how certain terms, phrases and even sentences had been translated in the past.

The growth in both the storage capability and processing power of personal computers in the 1980s eventually enabled the development of a number of commercial computer-aided translation tools inspired by the research done in the 1970s. Large databases of previous translations could be stored, indexed and searched efficiently, finally making the concept of TM available to ordinary translators.

As the use of TMs spread throughout the translation industry during the 1990s, technology vendors realized that TM technology could be marketed further up the chain of production, with the promise of a considerable reduction in translation costs (Gordon, 1996). While the initial focus of the research into translation aids had been primarily on extending the capabilities of the user, the emphasis soon shifted from increasing the efficiency of translators to reducing costs – first for LSPs (Language Service Providers) and eventually for translation buyers (LISA, 2004). This change is perhaps most clearly demonstrated by the emergence of de facto standard discount schemes based on the results of TM preprocessing: words that appear in segments which have a corresponding exact or fuzzy match in the TM are charged at lower rates.

This is not surprising, given that the commercial adoption of any new technology must be driven primarily by quantifiable measures of return on investment. However, the advent of the Internet, the resulting growth in the demand for translations, and the corresponding growth of the translation industry led, in turn, to an increase in the availability of TM vendors to such an extent that the technology became commoditized, and LSPs and translation buyers turned to workflow automation as the next area on which to focus their cost-cutting efforts.

The main obstacle in the road to fully exploiting the capabilities offered by Translation Memory is the fact that the translation industry is still focused on applying TM technology at the segment level, foregoing the advantages offered by treating translation databases as large parallel corpora. The most obvious reason for this is that special pricing schemes involving discounts for TM matches are only reasonable at the segment level, if at all. Any advantages obtained by applying TM technology at the subsegment level are therefore only of immediate value to the freelance translator, since they cannot be automated by the LSP. Additionally, given the nature of the translation tool market – where the specific software used by the freelance translator is often imposed by the LSP – freelancers have very little power as consumers when it comes to influencing the evolution of commercial TM tools. Together, these two idiosyncrasies of the translation industry have effectively resulted in a stagnation of the commercial research and development efforts towards improving and expanding the use of TM technology.

In spite of these commercial barriers, there are several developments that have appeared in recent years and which, if properly pursued, could lead to significant changes in the level of translation reuse afforded by Translation Memories.

Bitexts

While the use of parallel corpora as a translation reference tool can be traced back to 1988, when Brian Harris coined the term “bitext” (Harris, 1988), bitext-based translation tools only became widely known in the early 2000s, following their successful implementation at RALI (Recherche appliquée en linguistique informatique) (Macklovitch, Simard & Langlais, 2000) and commercial development by a number of Canadian vendors.

In their modern incarnation, bitexts have been promoted as an alternative to traditional segment-based TMs, since by their very nature as parallel corpora they preserve the context of the translations and, more importantly, they are also amenable to being queried not only for entire sentences or paragraphs, but also for smaller segments that might be useful.

The importance of context cannot be overstated – in traditional TM systems, the likelihood of finding different translations for the same source segment, particularly for short sentences, is surprisingly high. Even if the candidate TUs (translation units) are tagged with appropriate domain and/or client information, it may not be possible for the system to select the correct one automatically. The advantage provided by the bitext is that the context of the match can be instantly compared to the text being translated, thereby allowing the system to discriminate between the alternatives.

That said, the idea that traditional segment-based TM systems cannot implement such a feature – or, for that matter, any other use of contextual information – is manifestly mistaken. In recent years, most commercially available TM tools have introduced mechanisms which allow them to preserve the order to the TUs stored in their databases, thereby making it feasible to reconstruct the original context on-the-fly.

The only case in which bitexts might be superior is in the handling of incorrect segmentation of the source text, or instances where there is a many-to-one correspondence between sentences in the source and the translation. In such cases, the TM system may not find a match for a source segment, whereas a bitext could do so easily. Of course, even this could be overcome in a traditional TM system by automatically grouping contiguous sentences for which no match has been found in the TM, although it would be clearly much more computationally expensive that searching in a bitext.

On the other hand, bitexts have two serious drawbacks (Gow, 2003). Since they are not guaranteed to be correctly aligned, their use cannot be easily automated in the way that TMs are currently used. Additionally, since they are only generated once a translation has been completed, they cannot exploit the possibility of internal repetition within the source text, which limits their use for large, repetitive texts, particularly those that are worked on by teams of translators.

That said, the appearance of both research and commercial bitext-based translation tools has spurred the development of comparable functionality in modern segment-based TM systems and, in particular, has brought the concept of subsegment-level matching to the attention of most translation tool users.

Subsegment-level Matching

It should be readily apparent that, if we apply Translation Memories exclusively at the segment level, they are only going to be useful when dealing with certain types of highly repetitive texts, such as revisions of previous documents or documentation for new products that differ only slightly from previous models. This approach completely overlooks the repetition that may be present at the subsegment level, which is harder to take advantage of but can still be very useful for the translator (Simard & Langlais, 2000). Therefore, the next logical step towards making better use of TMs is to go beyond exact and fuzzy matches at the segment level, and use the TM as the reference tool it was originally envisaged to be.

Although the concept was popularized by bitext-based translation tools, subsegment-level lookup has been available for some time even in traditional TM systems (Melby, 2006), albeit in a rather rudimentary form. In most tools, subsegment-level lookup is implemented as a simple concordance search which displays all of the TUs in the TM which contain a specific “chunk”, usually a term or phrase. While such a simple feature may be immensely valuable as an extension of the translator’s own memory, its effectiveness may decrease when used on large TMs accumulated over a number of years; if a specific subsegment appears in hundreds of segments in the TM, the amount of time required by the translator to examine all of the available matches could make the feature counterproductive.

Viewing the TM as a large parallel corpus – or as a generator of virtual domain-specific parallel corpora – and applying basic statistical analysis techniques to it is one way in which subsegment matching could improve. Rather than simply displaying a large number of TUs containing a phrase, a TM system could analyze all of the translations for those segments and propose the most likely subsegment in the target TUs as the translation for the source subsegment (Simard & Langlais, 2000).

Implementing such a feature to provide real-time results poses several challenges: finding the longest common word subsequence(s) of a large set of segments is computationally expensive, so the translated segments must already be present in a form which simplifies such calculations. For example, systems which index all languages in a multilingual TM might be able to use the pre-existing indices for the translated segments to significantly improve the performance of these searches.

The second challenge involves dealing with inconsistent translations of the same phrase, either due to the fact that the translation varies depending on the context, or simply because there are several valid translations for it. If there is not enough material in the TM so that the frequencies of all the valid translations are statistically significant, the system may not be able to confidently identify any translation at all.

Finally, when trying to statistically determine the translation for a given phrase, it may not be possible to accurately detect the correct boundaries in the translation. If the translated phrase consistently appears together with a word that is not part of the phrase, such as a preposition or particle, the system may not be able to decide whether that word is part of the translation or not, particularly if there are not enough examples of the translation. The use of the Marker Hypothesis (Green, 1979) in many EBMT (Example-Based Machine Translation) implementations (Gough & Way, 2004) provides a reasonable solution, by using language-specific data on closed word classes which can be used to detect likely phrase boundaries.

Having solved these problems, candidate translations for subsegments found in the source text could be offered to the translator, who would then piece together a translation using the available chunks. One drawback to this approach, however, is that the amount of copy and paste operations and post-editing required to produce a valid translation may considerably reduce the potential productivity gains. On the other hand, offering those translations in the context of a predictive typing mechanism could eliminate most of that overhead (Simard & Langlais, 2000).

Fixing Fuzzy Matches

Notwithstanding the productivity gains that the individual translator can obtain from the proper application of subsegment-level matching in TM-based translation tools, such improvements in the technology are unlikely to become widespread in the translation industry unless they can be automated and translated into direct cost savings for LSPs or translation buyers, as we discussed above.

EBMT, particularly as originally envisioned (Nagao, 1984), can provide clues as to how these techniques could be integrated into non-interactive TM processes. The concept of translation by analogy could be applied to the reuse of past translations at the subsegment level by attempting to “repair” fuzzy matches retrieved from the TM.

The idea is simple: when the system encounters a source segment for which a TU found in the TM is only an approximate match, it can detect the differences between the segments and, using the same subsegment-level matching mechanisms outlined above, attempt to determine the translations for each of the differences, in order to perform the corresponding substitution in the translated segment.

While such a mechanism is not at all guaranteed to produce correct translations, it would certainly reduce the amount of post-editing required for a number of fuzzy matches, and would most probably lead to quantifiable potential savings which can be exploited by LSPs and translation buyers.

QA and Consistency

The subsegment-level matching mechanism outlined above can be considered a form of bilingual phrase alignment, a field in which a large amount of research has been done, both by SMT (Statistical Machine Translation) and EBMT researchers. While phrase alignment is an essential part of those two approaches to MT (Machine Translation), there are two areas where it could be useful for the translation industry in general: multilingual terminology extraction and consistency checking.

In fact, bilingual terminology extraction is already present in a number of commercial translation tools, although current implementations tend to require a fair amount of manual validation, and are very computationally intensive. Both these problems would be alleviated to a certain extent by performing terminology extraction on suitably preprocessed TMs, as explained above.

Consistency checking, on the other hand, has not yet been developed to such an extent. While most current tools have the ability to validate the translator’s work against termbases or glossaries, pointing out where specific terms have not been translated in the prescribed manner, the ability to identify phrases that have not been translated consistently (whether inside the same text, or as compared to existing TMs), would greatly increase the quality of translations.

Additionally, the same consistency checking procedures could be applied to existing TMs, thereby providing a measure of the quality of the data itself, and perhaps guidance as to how to best maintain them to reduce the amount of noise. While the increase in quality would be harder to quantify, particularly in terms of how it translates into return on investment, the decrease in support and litigation costs (Gow, 2003) should be enough to generate interest in these applications on the part of translation buyers, thereby ensuring their adoption by the translation industry at large.

Language Independence

Having established the existence of various marketable uses of advanced subsegment-level matching, we can consider what would seem to be the next logical step in the evolution of the technology: the application of specific linguistic knowledge to TM systems. This would ensure that the results of the fuzzy match repair mechanism described above are grammatically and syntactically correct, and that the handling of inflected terms in terminology lookups is improved.

To do so effectively would require the addition of morphological analyzers, POS (part of speech) taggers and full grammars for each of the languages supported by a system. In addition to the cost of developing or acquiring such linguistic resources, the question of coverage, particularly in corpus-based morphological analyzers and POS taggers, when dealing with unknown terms – something highly likely in technical translations – severely limits their use. The use of closed class word lists in conjunction with the Marker Hypothesis described above would, on the other hand, have a relatively low cost and should be seriously considered.

However, abandoning a purely statistical approach to TM technology would have a severe consequence: one of the main advantages of TM over MT, though rarely discussed, is its applicability to minority languages (Gow, 2003), since TM systems can be implemented without any additional linguistic information beyond what is provided by the NLS (National Language Support) and frameworks provided by the platforms on which they run.

Restricting the range of languages on which TM could be used would close off potential markets, and would ignore the fact that the main goal of TM is not to produce perfect translations automatically, but to increase the productivity of the human translator. If inflected terminology can be handled correctly using alternative approaches, and high-quality repaired fuzzy matches still need to be manually reviewed, the disadvantages outweigh the advantages.

Conclusion

It is clear that there is plenty of scope for TM technology to evolve and provide considerably higher levels of translation reuse as well as other complementary productivity increases. However, the main obstacle towards achieving this has nothing to do with the technical complexity of the proposed improvements, and more with the perceptions of the translation industry in general. As long as new advances in TM technology are only marketable in terms of time savings for translators, or of increases in quality and consistency that are not easily measured, it is unlikely that the translation buyers and LSPs who determine which translation tools are used throughout the market will adopt them. The development of new metrics to cover consistency and subsegment-level translation reuse will be the critical first step towards developing new approaches to marketing the next generation of TM technology.

Bibliography

Gordon, I. (1996). “Letting the CAT out of the bag - or was it MT?”, *Translating And The Computer 18: Proceedings of the Eighteenth International Conference on Translating and the Computer.* London: ASLIB.

Gough, N., & Way, A. (2004). “Robust Large-Scale EBMT with Marker-Based Segmentation”, *Proceedings of TMI 2004*, 95–104. Baltimore, Maryland.

Gow, F. (2003). *Metrics for Evaluating Translation Memory Software.* MA Thesis, University of Ottawa, Ottawa.

Green, T. (1979). “The necessity of syntax markers: Two experiments with artificial languages”, *Journal of Verbal Learning and Behavior*, 18, 481-496.

Harris, B. (1988). “Bi-text, a new concept in translation theory”, *Language Monthly*, 54, 8-11.

Hutchins, J. (1998). “The Origins of the Translator's Workstation”, *Machine Translation*, 13 (4), 287-307.

LISA. (2004). *2004 Translation Memory Survey.*

Macklovitch, E., Simard, M., & Langlais, P. (2000). “TransSearch: A Free Translation Memory on the World Wide Web”, *LREC 2000 Second International Conference on Language Resources and Evaluation.* Athens.

Melby, A. K. (2006). “MT+TM+QA: The Future is Ours”, *Tradumática*, 4.

Nagao, M. (1984). “A framework of a mechanical translation between Japanese and English by analogy principle”, in A. Elithorn, & R. Banerji (eds.), *Artifial and Human Intelligence.* Elsevier Science Publishers.

Simard, M., & Langlais, P. (2000). “Sub-sentential Exploitation of Translation Memories”, *LREC 2000 Second International Conference on Language Resources and Evaluation.* Athens.