[*Translating and the Computer 21*. Proceedings… 10-11 November 1999 (London: Aslib)]

A Building Blocks Approach to Translation Memory

Kevin McTait*, Maeve Olohan**, Arturo Trujillo*
Centre for Computational Linguistics*, Centre for Translation Studies**
Department of Language Engineering
UMIST
Manchester
PO BOX 88
M60 1QD

{kevinm,maeve,iat}@ccl.umist.ac.uk

Abstract: Traditional Translation Memory systems that find the best match between a SL input sentence and SL sentences in a database of previously translated sentences are not ideal. Studies in the cognitive processes underlying human translation reveal that translators very rarely process SL text at the level of the sentence. The units with which translators work are usually much smaller i.e. word, syntactic unit, clause or group of meaningful words. A building blocks approach (a term borrowed from the theoretical framework discussed in Lange et al (1997)), is advantageous in that it extracts fragments of text, from a traditional TM database, that more closely represent those with which a human translator works. The text fragments are combined with the intention of producing TL translations that are more accurate, thus requiring less post-editing on the part of the translator.

Introduction

Translation memory (TM) systems have proved extremely successful in a range of translation tasks, including the translation and updating of software manuals, revised texts, budgets and various other technical and highly repetitive materials. Current TM systems operate on the basis of a fuzzy matching mechanism that allows previously translated sentences or other text fragments to be retrieved and used as the basis for the translation of the current source language (SL) sentence.

One problem with such systems is that they cannot combine fragments from different translation units (TUs) to build up the target language (TL) translation presented to the translator. In general, TM systems can only handle a single TU at a time, and therefore their effectiveness relies on a fairly close match between the current sentence and a previously translated one. Admittedly, some newer TM systems can translate prices, product names, dates and other similar strings found in texts. Where this is possible, such strings can be translated very accurately, but for many kinds of texts, such additional functionality is of limited use.

In this paper, we argue that present TM systems are far from ideal, both from a practical perspective and from the point of view of emulating translation processes in humans. We describe an implemented system (McTait & Trujillo, 1999) which is capable of combining fragments from different TUs to produce a more accurate draft translation requiring less post-editing on the part of the translator. Our approach relies on extracting translation patterns consisting of non-contiguous text fragments occurring more than once in the database of TUs. During translation, fragments of the

input sentence are matched against fragments from the translation patterns. The result is a set of TL text fragments that are then combined to produce the TL string suggested to the translator.

First, we describe the two types of system that motivate our approach: TM and Example-Based Machine Translation (EBMT). Subsequently, we motivate our approach from a translator's perspective, and finally we describe our system in some detail.

Translation Memory

A typical TM system works by accepting SL input, typically sentences, from the translator and retrieving and ranking a set of closely matching TUs. From these, one or more is selected and the TL side is used as a draft translation. Ranking in a TM system is carried out by comparing the SL input sentence with the SL side of the TUs stored in a database. Selection of a TU from the ranked list is done either automatically, selecting the top ranked TU, or manually by the translator.

Conceptually, TM is a storage and retrieval mechanism that allows retrieval of partially matching items. It relies on two important operations for its effectiveness. First, it is necessary to measure how similar a sentence is to another in order to provide useful rankings. The ideal similarity measure would identify sentences whose translation is closest to that of the SL text being translated. In practice, this ideal score can only be approximated, for example by measuring the number of words in common between the input and the database sentences. It is also possible to use relative word positions, partial word matching and other textual clues to improve on the usefulness of the similarity scores.

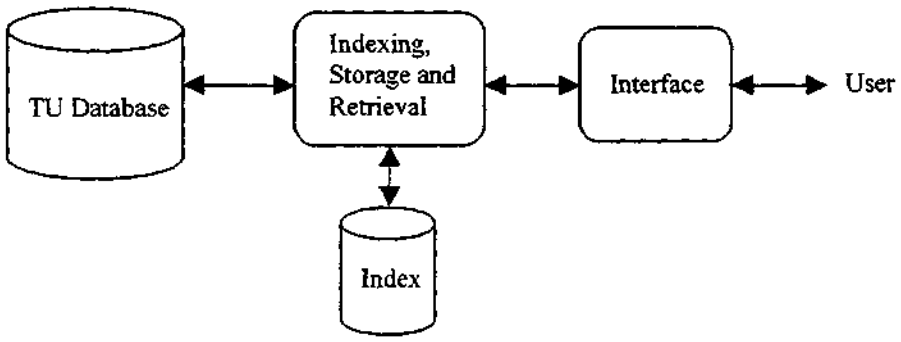

Figure 1: Structure of a TM System

Second, the speed with which TUs are retrieved determines the feasibility of TM in practical applications. Techniques such as inverted files and other indexing mechanisms have been used in information retrieval and are well understood. Figure 1 shows a typical TM system architecture that makes use of efficient indexing of TUs. An inverted file in a TM setting, for example, would indicate, for each word, the sentences in which it occurs. By looking up the input words in the inverted file, the sentences sharing the most words with the input can be efficiently identified and retrieved. Additional processing can refine this technique by taking into account

relative word order, morphological differences between words, high frequency and function words, and any other clue that can assist in determining sentence similarity[1].

As described above, TM is a relatively well understood problem, but it is restricted in that it can only deal with one TU at a time. Consider the following simple database of TUs. It is highly contrived, but serves to clearly illustrate the need for combining fragments from different TUs.

	English	Spanish
1	The program failed	El programa falló
2	The program worked	El programa funcionó
3	The computer failed	La computadora falló

Imagine now that the sentence to translate is:

The computer worked

None of the English sentences match this input fully, but, in some sense, all the relevant information is contained within the TM database. If one could use the subject of TU 3 and the verb of TU 2, a correct translation could be produced: *La computadora funcionó.* A TM system of the kind described above would fail to construct such a translation even though the repetitive nature of the database allows sub-sentential translations such as '*The program - El programa'* to be established with a high degree of certainty.

EBMT

The idea of combining fragments from existing translations to produce a new translation is not recent. Example-Based Machine Translation (EBMT) (Nagao 1984) is fundamentally dependent on this principle. It relies on a database of examples of previously translated sentences against which the SL input is matched. Contrary to TM systems, however, example fragments can match all or part of the input. The translations of these fragments are then combined to form the translation of the input.

Most formulations of EBMT assume that the translations of example fragments are identified in the example database. In addition, the combination of fragments to produce the TL sentence relies on a significant amount of structure being provided in the example database. Finally, matching examples are ranked based on a metric which includes a semantic classification of the words in the language.

A typical example database in EBMT contains grammatically analysed source and target sentences, together with the translation relations that exist between example fragments (Sato 1995). Figure 2 illustrates a sample entry in which different portions of the sentences are set in translation correspondence. Constructing such entries automatically requires significant linguistic resources, in the form of parsers and possibly bilingual dictionaries. Obviously, example entries can also be constructed manually but this requires some familiarity with the dependency analyses and involves additional intervention on the part of the user.

[1] The reader is referred to Trujillo (1999) for details on these and other relevant techniques.

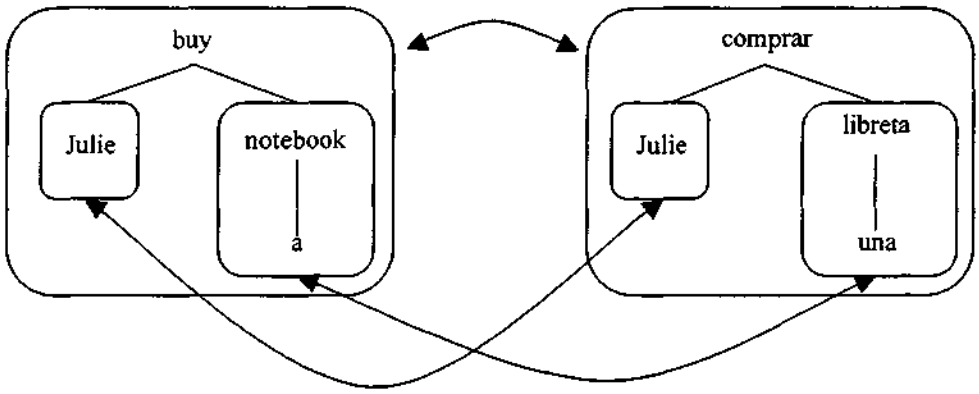

Figure 2: Correspondences in an EBMT Example Entry

Relying on large scale linguistic resources makes systems more expensive and difficult to acquire. Given the state of the art in computational linguistics, such resources are not guaranteed to produce correct results either. Finally, for some languages large scale resources are either not readily available or difficult to obtain. We therefore argue that neither TM nor EBMT offer a satisfactory solution to the Machine Aided Translation (MAT) problem.

TM and EBMT can be seen to lie at opposite ends of a spectrum in memory-based translation. On the one hand, TM requires few linguistics resources but cannot combine fragments from different TUs, and on the other hand, EBMT can combine example fragments, but does so by relying on parsers and other knowledge-intensive tools.

The novelty of our approach is that it does not rely on linguistic knowledge such as that found in POS-taggers, syntactic parsers, chunkers or similar knowledge-based tools to combine fragments from different TUs. The algorithm we have developed relies only on the information present in the TUs stored in a typical TM database or bilingual aligned corpus.

Motivation

Studies of human translation processes can provide us with a sound basis when designing and implementing a system which is capable of combining textual fragments. Investigations of a number of aspects of processing can be seen to support the approach taken here.

Empirical Translation Process Research

Prior to the growth of interest in empirical studies of the translation process, a number of theoretical models were developed[2]. Lorscher (1991: 7-27) critically evaluates

[2] See House & Blum-Kulka (1986) for some of the first papers dealing with the empirical investigation of the translation process, notably those written by Krings, Gerloff and Lörscher, who discuss

those developed by Diller & Kornelius (1978), Nida (1969) and Kade (1968) and that implemented by Hönig & Kussmaul (1982). Apart from various criticisms levelled at individual models in relation to particular aspects such as their definition of translation and their degree of generality or oversimplification, all these models are, in Lörscher's view, fundamentally flawed since they are either prescriptive, i.e. presenting what the translator should ideally do in translating, or they describe the translation process as static and therefore do not account for the interplay between various components of the process (Lörscher 1991: 26). Furthermore, they do not, in his view, 'account for the psychological reality of translating', i.e. it is not possible to draw conclusions from them about what mental processes and/or subprocesses occur in translation (op. cit: 26). This is a particularly significant shortcoming if one expects a model of the translation process to deliver or account for information pertaining to the cognitive processes involved in translation. Thirdly, the models presented are 'theoretical-speculative', developed by rational deduction rather than empirically (op. cit.: 27). Again this is a crucial issue in the search for knowledge about what real translators, whether professional or trainee, actually think and do when translating, as opposed to adherence to a notion of what should ideally happen in translation.

Introduction to Empirical Investigations of Translation Processes

Studies of cognitive processing in translation have been carried out using introspective methods borrowed from psychology, in particular the think-aloud method of data elicitation. This involves the translator verbalising his/her thoughts while translating. This output is recorded, transcribed to constitute a think-aloud protocol or TAP, and analysed. There are some limitations and shortcomings in terms of the methodology[3], and it is acknowledged that a complete picture of the contents and processes of the 'black box' cannot be produced in this way. However, this method remains a valid means of accessing something of the translator's thought processes, providing information about how translators approach the task, how they solve problems and make decisions. While some of the initial empirical investigations aimed to model the entire translation process and to identify and describe all translation strategies (e.g. Krings 1986), more recent studies have focused their investigation on specific aspects of the process, increasingly translator-centred e.g. looking at affective factors in translation; Tirkkonen-Condit & Laukkanen (1996), Jaaskelainen (1996), Tirkkonen-Condit (forthcoming), House (forthcoming)[4].

We will address two main areas in which translation process research findings are directly related to the design and implementation of a translation memory system which operates on the level of textual fragments: investigations into the unit of analysis or unit of translation and the progression of translators through the text, and studies of the automation or routinisation of processes.

---

translation as carried out by second language learners. Other papers in the same volume, namely those by Faerch & Kasper, and by Börsch clearly focus on second language learning rather than translation.

[3] For discussion of introspective methods, see Ericsson & Simon (1984).

[4] For a more comprehensive discussion of translation process research and its findings, see Olohan (1998).

Unit of Analysis and Progression through Text

The unit of processing constituted an important focus of attention in early studies of the translation process carried out by Gerloff, one of the pioneers in the field. In her pilot study, she examines second language learner processes of text analysis using a translation task, looking in particular at the cognitive operations underlying comprehension and production of language, and at the relationship that exists between these two processes (Gerloff 1986; 1987: 137). In using a translation task, Gerloff elicits information about retrieval strategies and strategies of analysis, editing and inference for purposes of both comprehension of the L2 and production of the L1 (Gerloff 1986: 244). She examines think-aloud protocols and translations produced by five students and identifies the unit of analysis, whether morpheme, word, phrase, clause, sentence, discourse or group (non-syntactic grouping of words) (Gerloff 1986). The units observed in the case of the students' translation are then compared to a TAP produced by a bilingual (Gerloff 1987) and, in doing so, differences are perceived in the units of analysis and in other characteristics of the translation process. These differences can be attributed to various levels of L2 proficiency and varied degrees of training in translation. Additional data are subsequently employed in a larger-scale analysis of the translation processes of four students, four bilinguals without professional translation experience, and four professional translators (Gerloff 1988).

In tracing patterns of movements through the text, again through the processing observable in the TAPs, she finds evidence of the 'highly retrospective-prospective nature of the translation process' (op. cit.: 114). This can be seen from backtrackings and non-linear forward movement through the text. This term is used initially by Lörscher who likens the translation process to a chain of spirals or a chain of loops (Lörscher 1986: 16). Gerloff supports this view and concludes that direct progression from A to B may not constitute maximum efficiency in translation, and that backtracking may help create a sense of context (Gerloff 1988: 130-131). Not only are translators working in this retrospective-prospective way, but they are also processing at various levels simultaneously (op. cit.: 131-132).

The TAPs reveal that all of the subjects involved process largely in naturally-occurring syntactic units. Translators and bilinguals have approximately the same amount of processing at discourse level (7% and 6% respectively), while students exhibit less processing at this level (Gerloff 1988: 116-117). Generally there is very little processing at sentence level. Word, phrase and clause unit processing are the most frequent types of all participants, with students and bilinguals processing more at word level than clause, and this relationship is reversed for translators, who also process more at clause level.

In comparing the performance and strategies of students with those of professional translators and identifying differences in translation strategies between language students and professional translators, Lörscher asserts that, similar to Gerloff's findings in relation to differences in units of analysis, the units of translation of students are much smaller than those of professional translators (Lörscher 1996). The unit of translation is defined as:

> 'the SL text segments which the subjects extract and put into their focus of attention in order to render them into the target-language *as a whole'* (op. cit.: 30 - Lörscher's emphasis).

It can be seen that the findings of Gerloff and Lörscher bear considerable similarity to one another in terms of the units in which processing, analysis and translation are carried out, and the differences in these units when processing of students is compared with that of professionals. These studies show that L2 competence influences the units in which analysis or translation of the L2 takes place. Greater linguistic competence will result in a higher proportion of processing occurring at the level of larger units. Processing predominantly at word level is an indication of sign/form-oriented translation, often performed by language learners. Processing partly at clause level and above is indicative of sense-oriented translation (Lörscher 1992: 413). These findings may be seen to have relevance for later studies which focus on local and global decision-making, identifying linguistic and non-linguistic criteria for decision-making[5], and studies which look at the balance between top-down and bottom-up processing[6]. The difference in sizes of processing units has an impact on the level on which decisions are taken and on the criteria used for making decisions. Furthermore, both researchers find evidence for the retrospective-prospective nature of the translation process, and show that think-aloud studies can furnish information about patterns of movement through the text and the task.

Automation of Processes

Königs's work (1987) has as its objective the development of a psycholinguistic theory of translation, linked to empirical investigations and leading to conclusions for the teaching of translation. He introduces the notions of 'adhoc block' and 'rest block' in relation to processing. In the adhoc block, a one-to-one equivalence is drawn by the translator between source text and target text units, and processing happens more or less automatically. The rest block contains all other processing, including translation problems, linguistic or content-related. All information about the translation situation and the application of specific translation techniques also belong to the rest block. With this model as a basis, Königs carries out his empirical investigation, identifying what is translated adhoc, what elements belong in the rest block and what happens there. He discovers that adhoc translations are to be found up to sentence level. Here, translations are produced spontaneously and without immediate correction, often drawing on associations or past experience. Elements of the rest block, on the other hand, prompt hesitation on the part of the subject. These hesitations are attributed to a variety of possible reasons: gaps in L2 competence, gaps in translation competence, specific linguistic-based translation difficulties on word, sentence or text level, specific translation difficulties based on the content, and performance difficulties.

Jääskeläinen & Tirkkonen-Condit (1991) collaborate to examine the processing of three fifth-year (described as professional) and four first-year (referred to as non-professional) translators in an attempt to gain insights into those aspects of translation that tend to become increasingly automated as the translator's professionalism increases. As translators re-encounter units of translation, their translation becomes

---

[5] See, for example, Tirkkonen-Condit (1992) and Jääskeläinen (1993).
[6] This is investigated and discussed in Kussmaul (1995).

more automated, since a direct linguistic mapping is made possible and the need to activate a conceptual representation is eliminated. This is the insight we attempt to model in our system, since the goal is to identify recurring textual units. Their investigation is conducted on the basis of protocols which had previously been used in analyses of various aspects of the translation process cf. Tirkkonen-Condit (1989); Jääskeläinen (1989a), (1989b), (1993). They identify where processing was cognitively controlled and thus verbalisable in the non-professionals' protocols and where the same processing appeared to be automated and therefore not verbalisable in the professionals' protocols. They distinguish between the automation of procedural knowledge and of linguistic knowledge. By comparing protocols, they find examples of automated processes in the professionals' protocols. These processes are described as local e.g. linguistic thematic restructuring. They subsequently turn their attention to global decisions e.g. those related to the translation task and the written task description. They find that, whereas the non-professionals behave more or less randomly in this regard, the professionals appear to make global decisions consciously at an early stage and follow up unconsciously at a local level e.g. on a stylistic level. They therefore conclude that the professionals' decision-making can be said to have been automated during task performance. They see this as being in opposition to Königs's rest block and adhoc-block (cf. Königs 1987), which are perceived as more static.

Thus, while the terminology used to describe controlled and uncontrolled processes has varied from researcher to researcher, translation process data, in keeping with an information processing model, has shown than some processes are automatic or become automated in professionals, perhaps even during a single translation task.

Implications for Translator Tools

If one of the aims in the production of translator tools is to facilitate the translation process as it occurs in human translation, then it is clear that input from translation process research is both relevant and useful. Investigation of units of processing show that translators tend to process the text, not in sentence-sized chunks but in syntactically meaningful smaller units. It can therefore be considered useful for a translation memory system to be able to process in a similar manner. The translator's progression through a text is not linear and this supports the notion of a system which can apply and reapply its 'knowledge', i.e. stored translation units, to the translation. Furthermore, human processing becomes automated as strategies become proceduralised or routinised. Translation memory systems may be seen to operate in this way, but they crucially need to be able to do so with units of translation considerably smaller than the sentence.

Overview of System

The first stage of a system that operates on the basis of textual units that more closely represent those with which the human translator works, is the extraction of, what we term, *translation patterns.* These are extracted from a traditional TM database or sentence-aligned corpus and represent generalisations of sentences that are translations of each other. The translation pattern given in (1) shows how a sentence

in English containing *give...up* may be translated by a sentence in Spanish containing *abandono.* The variables *X* and *Y* stand for a series of one or more words and are referred to as *slots* as ultimately they are to be filled by text fragments from other TUs.

(1) X_s gave Y_s up $\longleftrightarrow$ X_t abandonó Y_t

Translation patterns or translation templates are traditionally extracted in the related task of EBMT (Kaji et al. 1992, Watanabe 1993, Takeda 1996) with the use of tools for linguistic analysis. The algorithm described in McTait & Trujillo (1999) extracts translation patterns without the need for such tools. This makes the system amenable to analysing any (European) language pair. Furthermore, a TM database made up of generalised sentential translations also facilitates matching a SL input in the database, since matches take place on the level of the text fragment.

A graphical representation of our system architecture is given in figure 3. While traditional TM makes use of a database of sentences that are translations of each other, in a building blocks approach to TM such a database is the input to an algorithm whose output is a database of translation patterns. Given a SL input sentence to translate, relevant translation patterns are retrieved from this new database and combined to form the TL translation. The aim of this process is to produce TL strings that are closer to the actual translation, thus requiring less post-editing on the part of the translator. Translations are suggested to the translator, ranked in best first order, as is done in traditional TM.

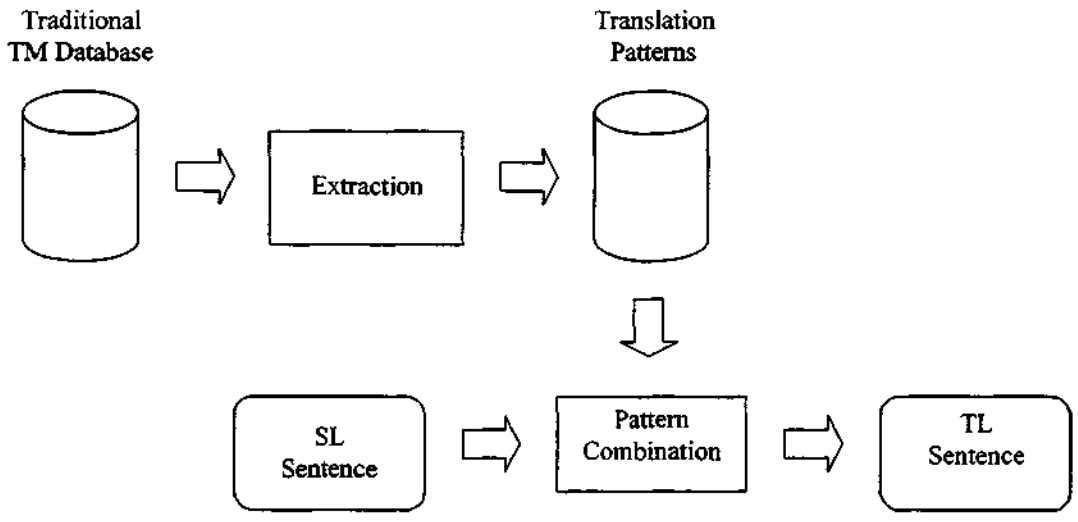

Figure 3: System Architecture

This new database and the corresponding step that combines translation patterns and text fragments from them represents the difference between a building blocks approach and a traditional approach to TM. In addition, once a new sentence has been translated by the translator, the database of translation patterns can be immediately

updated to include the information from this new pair of sentences, albeit in a slightly more complex process than that by which a traditional TM database is updated.

Translation Pattern Extraction

The algorithm to extract translation patterns operates on the principle that possibly discontinuous pairs of source and target strings that co-occur in 2 (or more) TUs are likely to be translations of each other. As an example, the translation pattern given in (1) is formed given the sample corpus in (2). Since *gave* and *up* appear together in both sentences and *abandonó* also occurs in both sentences, they are considered to be translations of each other. If the strings co-occur in many more than 2 sentences, the accuracy of the translation patterns increases (McTait & Trujillo, 1999).

(2) The commission gave the plan up ←→ La comisión **abandonó** el plan
Our government gave all laws up ←→ Nuestro gobierno **abandonó** todas las leyes

The text fragments and the slots (which essentially stand for text fragments also) are aligned, on the basis of their lengths in characters[7], to show the translation correspondences (as illustrated by the arrows in figure 4). Therefore, the set of alignments between text fragments in translation patterns can also be used as a bilingual lexicon or phrasicon.

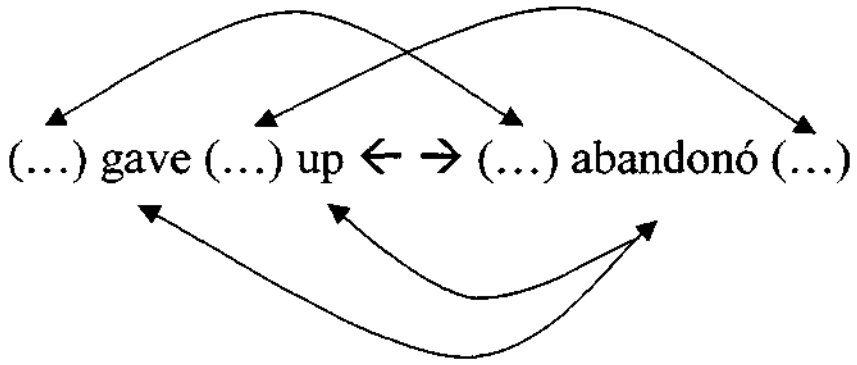

Figure 4: Translation Correspondences

Combining Translation Patterns

Given a SL input sentence, relevant translation patterns are retrieved from the database of translation patterns in a pattern-matching exercise. The longest and most similar translation patterns to the SL input sentence are retrieved. It is more than likely that these translation patterns are incomplete and thus contain slots which are to be filled by text fragments from other translation patterns in order to match the SL input sentence completely. Once enough translation patterns and text fragments are retrieved, they combine to form a TL translation string, using the alignments in the translation patterns themselves. This method is best illustrated through examples. The

[7] In a similar fashion to the way in which Gale & Church (1993) align sentences

following examples are taken from the translation patterns extracted using a 3000 sentence pair sample from the English/Spanish WHO AFI corpus[8].

Given the SL input: *aids control programme for Ethiopia,* the longest and most inclusive translation pattern from the database is retrieved (3). To complete the match with the SL input, a translation pattern that contains the text fragment *Ethiopia* is required (4).

(3) aids control programme for (...) ←→ programa contra el sida para (...)

(4) (...) Ethiopia ←→ (...) Etiopia

The formation of the TL string is then a simple matter of inserting *Ethiopia* into the relevant slot on the SL side of (3). On account of the fact that the slots in (3) are in translation correspondence and that *Ethiopia* and *Etiopia* are in translation correspondence, the translation of *Ethiopia* appears in the slot on the TL side of the translation pattern to produce the TL string *programa contra el sida para Etiopia.*

Of course, if a similar SL sentence is given e.g. *aids control programme for Thailand* then translation pattern (3) is still used and the database is searched for a translation pattern containing *Thailand,* and so on for other countries (5). Using this method, a more accurate translation can be formed, using information from more than one TU, consequently reducing the post-edit time for the translator.

(5) (...) Thailand ←→ (...) Tailandia
(...) Botswana ←→ (...) Botsuana
(...) the Caucasus ←→ (...) el Cáucaso

In order to fill the slots in a translation pattern, not all of the text fragments from another translation pattern are required, as in (3) and (4). If the SL input sentence is *humanitarian assistance to countries of former Yugoslavia,* translation pattern (6) is retrieved. In order to fill the slot in this pattern, translation pattern (7) is also retrieved. From (7) the text fragment *countries of former Yugoslavia* is required and inserted into the slot in (6). Since *countries of former Yugoslavia* is in translation correspondence with *los países de la antigua Yugoslavia,* and the slots in (6) are in translation correspondence, the TL translation becomes: *asistencia humanitaria a los países de la antigua Yugoslavia.* The fact that *assistance* and *asistencia* are aligned in (7) is, in this case, irrelevant.

(6) humanitarian assistance to (...)←→ asistencia humanitaria a (...)

(7) (...) assistance (...) countries of former Yugoslavia ←→
(...) asistencia (...) los países de la antigua Yugoslavia

Fuzzy Matching

Traditional TM systems make use of a fuzzy matching mechanism when there is no exact match between a SL input sentence and a SL sentence in the TM database. A

[8] The corpus is made up of headings and can be found at http://www.who.int/pll/cat/cat_resources.html

percentage score is often assigned to such matches, where only matches above a certain user-defined threshold are suggested to the translator. Fuzzy matching is also possible with our approach to TM. In a similar manner, matches are assigned a percentage score of similarity, again where only matches above a user-defined threshold are returned.

The building blocks approach to TM differs in that scoring takes place at the level of the text fragment or 'building block' as these form the fundamental units with which the system works (as opposed to the sentence). As an example, if the SL input sentence were *humanitarian assistance to countries of former Soviet Union,* translation patterns (6) and (7) could still be retrieved and combined as before. A measure of distance between the two text fragments is calculated in order to return a similarity score or percentage of similarity. This enables the system to decide whether the match is acceptable.

Linguistic Analysis

For a given SL input sentence, the translation patterns or text fragments may not always combine perfectly to produce a grammatical TL string. In some cases there will be a lack of agreement (number, gender etc) between constituents. For example, given the SL input *emergency assistance to war victims,* (8) and (9) might be the most similar translation patterns in the database. Combining the patterns would return the ill-formed TL string **asistencia de emergencia al víctimas de la guerra.* It is ill-formed since *al* is a contraction of *a el* where *el* is the masculine singular definite article in Spanish. What is required is *a los,* since *víctimas de la guerra* is plural.

(8) emergency assistance to (...)←→ asistencia de emergencia al (...)
(9)(...) war victims (...)←→ (...) víctimas de la guerra (...)

In other cases, words from the SL input would match more words in the translation patterns if their morphological differences were neutralised. It is desirable to be able to match inflectional variants of a lemma with each other and be able to say that *plays, played,* and *playing* are all forms of the same verb. This would increase the amount of patterns that could be retrieved for any given SL input sentence, which in turn would increase the chance of producing an accurate translation.

In an effort to solve these problems, we suggest an amount of linguistic analysis be incorporated into the system. We propose lemmatising the corpus in order to identify all inflectional forms of a lemma. This would allow the extraction of more general translation patterns. To ensure that there is agreement in the TL string, we suggest morphologically analysing the SL input and transferring that information to the translation patterns retrieved. Of course, introducing such language specific tools will make the system dependent on the resources available. However, morphological analysers/lemmatisers are perhaps the most commonly available and reliable analysis tools, and are widely available for a number of languages. We hope to report on the results of this work at a later date.

Conclusion

Our method of extracting translation patterns is a relatively simple one and has been proposed by Güvenir & Cicekli (1998). However, we believe our work differs from theirs in a number of ways. First, they view the problem as finding similarities and differences between two TUs, instead of using string co-occurrence. Secondly, in order to align the slots (what they term *differences)* they need to find those fragments of text elsewhere in the translation patterns they have extracted. This does not guarantee that every slot is aligned. In our approach, all slots are aligned (on the basis of lengths or cognates, for example). Thirdly, we also align the text fragments within the translation patterns themselves and not just the slots or differences. Fourthly, we can easily vary the threshold of string co-occurrence which in turn increases the amount of *differences* that we have to align. The result of this is that our approach is more robust and can produce more translation patterns and aligned text fragments. This means that there is a greater chance of covering a SL input and producing an accurate translation. Finally, to make our approach amenable to TM, we have also included the concept of fuzzy matches between text fragments, by using a measure of distance between two strings, to allow for greater coverage.

The building blocks approach to TM is advantageous in that it can produce translations by using information from more than one TU. The text fragments extracted from various TUs represent more closely the units of translation with which a human translator works. The principles of EBMT have been implemented in a language-neutral method. However, we propose to introduce a degree of analysis in the hope of producing more accurate translations. Furthermore, all the functionality of traditional TM (user-defined fuzzy matching thresholds, immediate inclusion of new translation pairs into the database etc) is included in a building blocks approach.

References

Diller. H. J. & J. Kornelius: 1978, *Linguistische Probleme der Übersetzung,* Tübingen: Max Niemeyer.

Ericsson, K. A. & H. A. Simon: 1984, *Protocol Analysis: Verbal Reports as Data,* Cambridge, MA.: MIT Press.

Gale, W.A. & K. W. Church: 1993, 'A Program for Aligning Sentences in Bilingual Corpora', in *Computational Linguistics,* 19: pp. 75-102.

Gerloff, P. 1986, 'Second Language Learners' Reports on the Interpretive Process: Talk-aloud Protocols of Translation', in House & Blum-Kulka (1986) pp. 243-262.

Gerloff, P. 1987, 'Identifying the Unit of Analysis in Translation', in Faerch & Kasper (eds.) (1987) *Introspection in Second Language Research,* Clevedon: Multilingual Matters, pp. 135-158.

Gerloff, P. 1988, *From French to English: A Look at the Translation Process in Students, Bilinguals, and Professional Translators,* Ann Arbor: UMI Dissertation Services.

Güvenir, H. A. & I. Cicekli: 1998, 'Learning Translation Templates from Examples', in *Information Systems,* 23: pp. 353-363.

Hönig, H.G. & P. Kussmaul: 1982, *Strategie der Übersetzung: ein Lehr- und Arbeitsbuch,* Tubingen: Narr.

House, J. (forthcoming). 'Consciousness and the Strategic Use of Aids in Translation', paper presented at AILA 1996.

House, J. & S. Blum-Kulka (eds.). 1986, *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies,* Tübingen: Narr.

Jääskeläinen, R. 1989a, Translation Assignment in Professional vs. Non-professional Translation: A Think-Aloud Protocol Study', in Seguinot (1989) pp. 87-98.

Jääskeläinen, R. 1989b, The Role of Reference Material in Professional vs. Non-Professional Translation: A Think-aloud Protocol Study', in Tirkkonen-Condit, S. & S. Condit (eds.) (1989) *Empirical Studies in Translation and Linguistics,* Joensuu: University of Joensuu, pp. 175-200.

Jääskeläinen, R. 1993, 'Investigating Translation Strategies', in Tirkkonen-Condit, S. & J. Laffling (eds.) (1993) *Recent Trends in Empirical Translation Research,* Joensuu: University of Joensuu, pp. 99-120.

Jääskeläinen, R. 1996, 'Hard Work will Bear Beautiful Fruit: A Comparison of Two Think-Aloud Protocol Studies', *Meta,* XLI/1: pp. 60-74.

Jääskeläinen, R. & S. Tirkkonen-Condit. 1991, 'Automatised Processes in Professional vs. Non-Professional Translation: A Think-aloud Protocol Study', in Tirkkonen-Condit, S. (ed.) (1991) *Empirical Research in Translation and Intercultural Studies,* Tübingen: Narr, pp. 89-109.

Kade, O. 1968, *Zufall und Gesetzmäßigkeit in der Übersetzung,* Leipzig: VEB Verlag Enzyklopädie.

Kaji, H., Y. Kida & Y. Morimoto. 1992, 'Learning Translation Templates', in *Proceedings of the 15th International Conference on Computational Linguistics:* COLING-92, Nantes, France, pp. 672-678.

Königs, F. G. 1987, 'Was beim Übersetzen passiert: Theoretische Aspekte, empirische Befunde und praktische Konsequenzen', *Die Neueren Sprachen,* 86/2, pp. 162-185.

Krings, H. P. 1986, *Was in den Köpfen des Übersetzers vorgeht: Eine empirische Untersuchung zur Struktur des Übersetzungsprozesses an fortgeschrittenen Französischlernern,* Tübingen: Narr.

Kussmaul, P. 1995, *Training the Translator,* Amsterdam: John Benjamins.

Langé, J-M, E. Gaussier & B. Daille. 1997, 'Bricks and Skeletons: Some Ideas for the near Future of MAHT', in *Machine Translation* 12: pp. 39-51.

Lörscher, W. 1986, 'Linguistic Aspects of Translation Processes: Towards an Analysis of Translation Performance', in House & Blum-Kulka (1986) pp. 277-292.

Lörscher, W. 1991, *Translation Performance, Translation Process, and Translation Strategies: A Psycholinguistic Investigation,* Tübingen: Narr.

Lörscher, W. 1992, 'Form- and Sense-oriented Approaches to Translation', in Thelen & Lewandowska-Tomaszczyk (1992) pp. 403-414.

Lörscher, W. 1996, 'A Psycholinguistic Analysis of Translation Processes', *Meta,* XLI/1: 26-32.

McTait, K. & A. Trujillo. 1999, 'A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns', in *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-99,* Chester, UK, pp. 98-108.

Nagao, M. 1984 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in A. Elithorn and R. Banerji (Eds.) *Artificial Intelligence and Human Intelligence,* pp. 173-80. Amsterdam: North-Holland.

Nida, E.A. 1969, Science of Translation, *Language* 45: pp. 483-498.

Olohan, M. 1998, *The Role of Theory in Enhancing the Explanatory and Predictive Aspects of Translation Process Research.* PhD Thesis, UMIST.

Sato, S. 1995, 'MBT2: A Method for Combining Fragments of Examples in Example-Based Translation', in *Artificial Intelligence* 75(1), pp. 31-49.

Séguinot, C. (ed.) 1989, *The Translation Process*, Toronto: H.G. Publications.

Takeda, K. 1996, 'Pattern-Based Context-Free Grammars for Machine Translation', in *Proceedings of the 34th Meeting of the Association for Computational Linguistics,* Santa Cruz, California, pp. 144-151.

Thelen, M. & B. Lewandowska-Tomaszczyk. 1992, (eds.) *Translation and Meaning:* Part 2, Maastricht: Rijkshogeschool Maastricht.

Tirkkonen-Condit, S. 1989, 'Professional vs. Non-professional Translation: A Think-Aloud Protocol Study', in Séguinot (1989) pp. 73-85.

Tirkkonen-Condit, S. 1992, 'The Interaction of World Knowledge and Linguistic Knowledge in the Processes of Translation. A Think-Aloud Protocol Study', in Thelen & Lewandowska-Tomaszczyk (1992) pp. 433-440.

Tirkkonen-Condit, S. (forthcoming) 'Uncertainty in Translation Processes', paper presented at AILA 1996.

Tirkkonen-Condit, S. & J. Laukkanen. 1996, 'Evaluation - A Key Towards Understanding the Affective Dimension of Translational Decisions', *Meta,* XLI/1: pp. 45-59.

Trujillo, A. 1999, *Translation Engines: Techniques for Machine Translation.* London: Springer-Verlag.

Watanabe, H. 1993, 'A Method for Extracting Translation Patterns from Translation Examples', *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-93: MT in the next Generation,* Kyoto, Japan, pp. 292-301.