

# *Example-Based Machine Translation: A New Paradigm*

Chunyu Kit, Haihua Pan and Jonathan J. Webster  
*Department of Chinese, Translation and Linguistics  
City University of Hong Kong*

## **Introduction - Why EBMT?**

Machine translation (MT) is aimed to enable a computer to transfer natural language utterances in either text or speech from one language into another while preserving the meaning and interpretation. MT technology has gone through several paradigms from its very beginning in the past half century, including word-to-word *direct* translation, rule-based *transfer* approach, *inter-lingua* approach and *knowledge-based* machine translation (KBMT). The direct translation approach relies too much on dictionary look-up. The transfer approach incorporates language analysis and representation at various linguistic levels, but cannot find adequate knowledge to resolve ambiguities involved in the language analysis, transfer and generation. The inter-lingua approach rests on the assumption that all languages share a common underlying representation, but such a goal appears unreachable. The *knowledge-based* approach attempts, mostly in the fashion of knowledge engineering (KE) in traditional symbolic AI, to acquire and encode various kinds of knowledge (e.g., encyclopedic knowledge) for the purpose of disambiguation, but the source of knowledge remains a serious problem. Most recent MT research falls into statistical MT and example-based MT. It is argued in Somers (1998; 2000a) that statistical MT is also an example-based approach.

In general, the translation process in traditional MT involves three sub-tasks, namely, analysis, transfer and generation, as depicted in the MT pyramid. Analysis deals with the transformation of the source utterances into a pre-defined format of internal representation, through morphological processing, part-of-speech (POS) tagging, syntactic parsing, semantic analysis, etc. Transfer works to convert the representation of the source language into that of the target language. Except for the inter-lingua approach, there used to be different representations for different languages. Generation, or synthesis, is concerned with the derivation of target utterances from the representation, observing necessary syntactic, semantic and pragmatic constraints. The generation process could be thought of as the reverse process of the analysis, but actually is quite different in nature. Each step in these processes involves, inevitably, many ambiguities, and each kind of ambiguity needs a large volume of knowledge for comprehensive disambiguation. For example, a bilingual dictionary is far less enough for word sense disambiguation (WSD) to determine what a word – which usually has many senses – exactly means within a particular context.

As in other language engineering tasks, MT needs a vast amount of knowledge for disambiguation, which is a kind of decision-making during the above translation processes. How to acquire adequate knowledge for reliable disambiguation is one of the most critical issues in current MT technology. Symbolic KE approach may have been a possible solution, if it had matured enough to successfully handle several significant problems in large-scale practical KE. For a practical language engineering task such as MT, the problems include how to acquire enough knowledge, how to represent and encode different kinds of knowledge, how to maintain the knowledge base and resolve the scale-up problem, etc. In the past decades it was the practice have experts (e.g., syntacticians) write down expert knowledge (e.g., grammar rules) in some rule-based format. This achieved significant but limited success. The inadequacy of manually encoded knowledge remains a problem – there are always so many practical ambiguities that experts can not foresee during the construction of knowledge base. The maintenance and scale-up problems also emerge when the knowledge base becomes larger and larger. For example, changing a single rule might cause unpredictable conflicts with other rules and, consequently, lead to a crash of the entire rule system.

Thus, there is a necessity to look for an alternative approach to knowledge engineering for MT that can automatically acquire practical knowl-

edge from, and also adapt itself towards, real language data. EBMT is considered one of the current attempts towards this goal.

The basis for EBMT is the existence of a large volume of translated texts (i.e., parallel bilingual texts), which have been translated by professionals with not only language proficiency but also specialist expertise. In this sense, bilingual texts encode knowledge that can be extracted to facilitate the automatic translation. Technically speaking, EBMT is about how to “decode” knowledge from bilingual texts, where the knowledge seems to have no overt formal representation or any encoding scheme. Instead, such knowledge is encoded in a way as straightforwardly as text coupling: a piece of text in one language matches a piece of text in another language.

In this article, we will give an overview of the EBMT technology. In the next section we will review the history of EBMT, with a focus on the main ideas. Since a comprehensive review of EBMT can be found in Somers (2000a), we will focus on the discussion of our viewpoints of the EBMT framework. Then we will define the notion of example and examine the major issues involved in EBMT, covering mainly the four major stages of EBMT, namely, example acquisition, example base management, example application and target sentence generation. Some of our current work in lexical-based text alignment for example acquisition is also discussed, highlighting the formulation of a similarity measure and alignment algorithm, before concluding our discussion in the last section.

## History

EBMT was most notably attributed to Nagao and his famous “translation by analogy” paper in 1984. Relevant ideas, however, can be dated back to 70’s, if not earlier. In view of this, it appears more appropriate to think of EBMT as having multiple origins, among which Kay’s (1997) *translation memory* (TM) and Nagao’s (1984)<sup>1</sup> *translation by analogy*, in our viewpoint, are the most influential ones.

According to Melby (1995), the Brigham Young University MT group incorporated a similar idea into the ALPS system, one of the earliest commercial MT systems, as a “Repetition Processing” tool. Arthen (1978) proposed “a programme which would enable the word processor to ‘remember’ whether any part of a new text typed into it had already been translated, and

---

<sup>1</sup>It was first presented in a 1981 conference (Somers, 1998).

to fetch this part together with the translation". He foresaw that a translator could benefit greatly from on-line access to similar translated texts while translating a new text, if "the system would check this text against the earlier texts stored together with other ... languages" and provide a simple operation as "cut and stick". Kay (1980) proposed the *translator's amanuensis*, i.e., a bilingual text editor as a workbench for translation, as a less ambitious but more realistic starting point for the MT technology, and put forward the idea of having the machine memorize existing translations to facilitate ongoing translation in his famous "proper place" paper on the relation of men and machines in language translation: "... the translator ... issuing a command causing the system to display anything in the store that might be relevant ... he can examine past and future fragments of text that contain similar material". This idea, having been suggested by others since the early 1970's (Hutchins, 1998), was later termed as TM. In general, TM can be understood as a restricted form of EBMT.

The essence of Nagao's EBMT, or "machine translation by the analogy principle", is more comprehensive, however, than TM. It attempts to mimic the cognitive processes of human translators for the purpose of automating the translation process. There are three main tasks in EBMT, as identified by Nagao, including

- Matching fragments against existing examples,
- Identifying the corresponding translation fragments, and then
- Recombining them to give the target text.

Among these three, TM shares the first. In Nagao's own words: "Man does not translate ... by doing deep linguistic analysis. Rather, man does translation, first by properly decomposing an input sentence into certain fragmental phrases ..., then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence." Nagao (1984) also proposed a matching technique based on measuring the semantic proximity of words. A similar idea was proposed independently by the DLT group in Utrecht at about the same time, according to Pappegaaij *et al.* (1986) and Schubert (1986).

In the early 90's, many MT researchers were attracted to two new paradigms of MT, one being statistical-based machine translation (Brown *et al.* 1990, Brown *et al.* 1993) and the other being Nagao's EBMT. In a review paper on EBMT, Somers (1998) notes that "The statistical approach is clearly example-based in that it depends on a bilingual corpus,

but the matching and recombination stages that characterise EBMT are implemented in quite a different way in these approaches.”

The EBMT approach became popular soon after some positive results were published in a number of papers demonstrating its plausibility. Sato and Nagao (1990) investigated the problem of example selection by approximate (or inexact) matching of input sentences and example sentences, using a similarity measure based on the syntactic similarity of dependency tree structures of a sentence pair in question and on the word distance (i.e., a similarity value) of corresponding words, which were pre-defined in a thesaurus. Sumita *et al.* (1990) looked into example-based translation of Japanese noun phrases of the pattern [N1 *no* N2] into English as [N2 *prep* N1] or [N1 N2], based on a distance measure for the input phrase and example phrase, calculated as a linear weighted sum of the distances of the three sub-parts, each of which is pre-defined in a thesaurus. However, the serious difficulty in constructing a large thesaurus with reliable similarity values between so many word pairs would prevent these two theoretically interesting approaches from having any practical application except for very limited domains. The problem of sparse data also looks like to prevent these two approaches, one limited to using sentence level examples and the other to phrase examples, from further success in practice.

By around 1993, EBMT had become an established research field of MT and many example-based techniques were applied to various MT tasks. Sato (1993) attempted the example-based translation of computer technical terms with respect to the focus term and its surrounding contexts and reported an overall accuracy of 96%, with an accuracy of 92% for unknown terms. Sumita *et al.* (1993) tackled the prepositional phrase attachment problem in translation with example-based techniques, using the same similarity measure as in Sumita *et al.* (1990). The research focused on nine most frequently used English prepositions and the positive results suggest that this approach can be generalized to many other prepositions and other types of structural ambiguity. Nirenburg *et al.* (1993) takes a closer look into the problem of matching an input sentence against possible examples, based on a distance measure defined in terms of necessary keystrokes in editing operations (e.g., deletion = 3 strokes, substitution = 3 strokes) required to convert an example into the input sentence. The experiments using WSJ data suggest that example matching at the sentence level is undesirable and partitioning a sentence into sub-sentential strings is necessary, in order to achieve an acceptable example matching result. Nirenburg *et al.*

(1994) and Brown (1996) reported their progress in integrating EBMT with KBMT.

Unfortunately, however, there has been little research on EBMT related to Chinese ever reported, although a number of researchers have explored the lexical acquisition at word and phrase levels by text alignment techniques. For example, Wu and Xia (1995) report successful results from English-Chinese text alignment at various levels using statistical-based methods; Fung and McKeown (1997) carried out English-Chinese terminology translation using a noisy parallel corpus.

### **What Is an Example?**

In the early research of EBMT, e.g., in early 1990's, many researchers tended to focus on examples at the sentence level. Since sentential examples that can exactly match input sentences are rare, researchers consider the possibilities of looking for similar examples or translation templates for translating new sentences. But the effectiveness seems problematic.

In general, an example is a pair (or couple) of texts in two languages that are a translation of each other. The texts can be of any size at any linguistic level: words, phrase, clause, sentence, and event paragraph. More flexibly, an example need not match a linguistically meaningful structure or constituent, that is, an example can simply be a pair of text chunks of an arbitrary length. However, we know that longer examples have a lower chance to show up in incoming texts. It is not strange that many sentential examples have no chance to be hit again during the phase of example application in EBMT. Thus, we need to pay attention to the usefulness of an example when considering what examples need to be acquired and put in the example base (EB) – the collection of examples.

A critical issue that needs to be examined closely in this context is the number of examples over a large-scale bilingual corpus, which can be unlimited in practice. Notice that an example can be further decomposed, in more than one possible way, into sub-structures or shorter examples, and that examples can overlap with each other. Therefore, the example number can be exponentially large in respect to the corpus size, if we collect all possible examples from a bilingual corpus. Consequently, the impracticality and implausibility of EBMT might arise, because any fragment of a sentence can be an example. We know that a language is well-known

for utilizing limited resources (e.g., lexicon and grammar) to produce an unlimited number of utterances. Thus, it is an interesting issue to examine the practicality of EBMT in terms of the correlation of example number and corpus size. In practice, how to control an EB to a reasonable size becomes vitally critical. For this purpose, we need to determine what examples should be filtered out and which ones should be maintained in the EB, not only for the matter of efficiency but, more importantly, for practicality. Although we have not come up with a clear strategy for example control, it is nevertheless an important issue in example base management.

The relation of bilingual dictionary and EB is also worth of careful examination. Conventionally, a bilingual dictionary is a collection of lexical entries in one language and gives many possible translations in another language for each word. We can think of a bilingual dictionary as a restricted EB, containing a collection of examples restricted at the word level. In return, an EB can be regarded as an extended bilingual dictionary. One might point out the fact that translating a word into another language following a bilingual dictionary is so uncertain, but translating a multiple word fragment of utterance in terms of an example from the EB is, in contrast, more sure. But this does not necessarily mean the dictionary and the EB do not share an intrinsic property for translation, namely, they provide choice of translation for a fragment, either single or multiple word, in an utterance. We can conceive - actually, have observed - that just like lexical entries in a dictionary, examples in the EB are also not limited to one-to-one mapping, because many utterance fragments can have more than one choice of translation. All choices need to be collected in the EB, highly similar to that in a dictionary, although the choices are significantly fewer.

Hence, an empirical MT approach like EBMT can be understood as to tackle the following problem: given some observed translation as a set of utterance fragments (either word, phrases, sentences or even some text chunks) with their possible choices of translation in another language, find a reasonably good, if not the best, translation for next input utterance.

## **The Four Stages of EBMT**

In general, there are four stages of work in EBMT, namely, example acquisition, example base management, example application and target sentence synthesis. Example acquisition is about how to acquire examples from par-

allel bilingual corpus (i.e., existing translation), and example base management is about how examples are stored and maintained. The example application concerns itself with how examples are used to facilitate translation, which involves the decomposition of an input sentence into examples and the conversion of source texts into target texts in terms of existing translation. The sentence synthesis is to compose a target sentence by putting the converted examples into a smoothly readable order, aiming at enhancing the readability of the target sentence after conversion.

### **Example Acquisition**

There are various resources from which we can collect examples. For example, from bilingual dictionaries we can collect examples at the word level. It is in this sense that an example base can be thought of as an extended bilingual dictionary that covers examples beyond the word level. These multiple-word examples have to be acquired from bilingual corpora, most of which are usually aligned at the article or even the paragraph level.

Text alignment seems to be a necessary step towards example acquisition at various levels. Manual alignment by experts can, of course, produce quite reliable examples, but the price for precision is a problem, and the speed is far less enough to handle a corpus of millions or tens of millions of words for practical applications. Thus, we have to resort to automatic text alignment technology.

The approaches to text alignment can be categorized into two types, namely, resource-poor and resource-rich approaches. The resource-poor approach mostly focuses on sentence alignment and relies mainly on sentence length statistics, co-occurrence statistics and some limited lexical information, as illustrated in Kay & Roscheisen (1988, 1993), Gale & Church (1991), Brown *et al.* (1991), Chen (1993), Church (1993), among many others. In addition to a collection of examples, a bilingual lexicon is also expected to be inferred from a parallel corpus via alignment, known as word alignment. In contrast, the resource-rich approaches make use of whatever available and useful, in particular, bilingual lexicon and glossary, to facilitate the alignment. We will have some more discussion on these approaches later.

Examples to be learned are not limited to the sentence level. Rather, we are more interested in learning examples at sub-sentential levels, including words, idioms and collocations, multi-word terminology, and phrases. The

alignment at the levels of phrase, collocation and word is more critical, because they are the examples whose target language parts are so productive in the composition of the translation output. Some critical techniques have been developed in previous research, e.g., word alignment in Dagan *et al.* (1993), partial parsing for EBMT in Furuse & Iida (1994), and bilingual parsing with the stochastic inversion transduction grammars in Wu (1995a, 1995b, 1997).

A comprehensive review on text alignment technology can be found in Wu (2000). The discussion below will focus on the work of text alignment with lexical resources that we are undertaking for the purpose of example acquisition.

### Similarity Measure

A similarity measure is required in text alignment to give an indication to which pair of texts in a bilingual corpus is more likely to be a match. For example, given a bilingual corpus that has an average sentence length ratio  $r$ , it is reasonable to believe that a pair of sentences whose length ratio is close to this  $r$  are more likely to match each other than two sentences whose length ratio is far different. The observation underlying this belief is that a short sentence in one language is usually translated into a short sentence in another language, and so are the long sentences. Also, sentence position is also useful information to exploit. Sentence pairs whose two sentences are in positions far away from each other in a parallel bilingual corpus have no doubt a lower chance to match each other than the ones that are closer to each other.

In addition to the factors of sentence length and position, a resource-rich approach may also take into account the matched pairs of dictionary items and glossary (or terminological) items. Accordingly, a formula for scoring the similarity of a sentence (or clause) pair  $\langle s_p, s_q \rangle$  in a given bilingual corpus can be empirically formulated as below:

$$sim(s_p, s_q) = \frac{\sum_{i=1}^m f(d_i) + w \cdot \sum_{j=1}^n f(g_j)}{|r \cdot l(s_p) - l(s_q)| \cdot |i - j|} \quad (1)$$

where  $p$  and  $q$  are sentence (or clause) positions,  $d_i$  and  $g_j$  are matched dictionary and glossary items, respectively,  $f(\cdot)$  is an evaluation function for the significance of a matched item in the similarity measure,  $w$  is a weight

indicating how many times a matched glossary item is as important as a matched dictionary item in the similarity measure, and  $l(\cdot)$  is the length of a given text. The simplest evaluation function is  $f(\cdot) = l(\cdot)$ , which means that we take the length, say, in characters, of a matched item as the measure for its significance in the similarity measure. Since a pair of matched items involves two parts in two different languages, the way to combine the length of the two into one looks less straightforward than simply summing up their lengths. An equivalence ratio of string length in characters in different languages needs to be taken into account. However, to simplify the issue, it would not be a poor choice to consider only the string length in one language.

The coefficient  $w$  is to be determined for individual corpus through experiments. According to our experimental results on BLIS, the English-Chinese bilingual corpus of HK legal texts of about 20 million words, the optimal value for  $w$  is 8, which indicates that a matched glossary item is about 8 times as important as a matched dictionary item in general.

Also, since matched glossary and dictionary items both give an indication of a good alignment overwhelmingly stronger than the clause length and position information, we have experimental results to show that omitting these two factors in the above similarity measure leads to no significant difference in our resource-rich approach to clause alignment for the parallel corpus of HK legal texts.

### Alignment Algorithm

When a similarity measure is available, we can calculate the similarity for all sentence pairs in a given bilingual corpus. This calculation produces a similarity matrix. For example, below is an illustration with a few sentence pairs, where  $sim(c_2, e_1) = 0.6$ . For simplicity, we may denote a score  $sim(c_i, e_j)$  as  $a_{ij}$ ; accordingly,  $a_{11} = 1.2$  and  $a_{21} = 0.6$ .

	$e_1$	$e_2$	$e_3$	$\dots$
$c_1$	1.2	2.3	0.4	$\dots$
$c_2$	0.6	0.9	2.5	$\dots$
$c_3$	0.7	0.8	7.5	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

Once a similarity matrix is available, the alignment algorithm to pick up a set of scores in the matrix that covers all sentences in the corpus can be

rather straightforward. Remember that the rule of thumb for the alignment algorithm is that every sentence in each language tends to match a sentence in another language with the highest similarity score. Following this, we can have an alignment algorithm as follows with regard to a given similarity matrix:

- (1) Pick the greatest score in each row;
- (2) Pick the greatest score in each column;
- (3) Derive the union of the two sets obtained from (1) and (2).

With the above matrix as an example, the algorithm picks  $\{a_{12} = 2.3, a_{23} = 2.5, a_{33}\}$  in step (1) and  $\{a_{11} = 1.2, a_{12} = 2.3, a_{33} = 7.5\}$  in step (2), and the union of the two sets is  $\{a_{11} = 1.2, a_{12} = 2.3, a_{23} = 2.5, a_{33} = 7.5\}$ , as marked in bold face below. Consequently, the alignment gives the result  $\{<[c_1] : [e_1, e_2]>, <[c_2, c_3] : [e_3]>\}$ , with the former maps one sentence (or clause) to two and the latter two to one.

	$e_1$	$e_2$	$e_3$	$\dots$
$c_1$	<b>1.2</b>	<b>2.3</b>	0.4	$\dots$
$c_2$	0.6	0.9	<b>2.5</b>	$\dots$
$c_3$	0.7	0.8	<b>7.5</b>	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

We can see that this simple algorithm derives not only one-to-one but also one-to-many alignments, and the many-to-many alignments are also conceivable. In practice, however, the interpretation of the set of scores chosen from a similarity matrix is not as straightforward as the above example, because not a few many-to-many alignments may be involved, although an overwhelming number of alignments are one-to-one. According to our observation, the many-to-many alignments are not rare in practice. An inadequate capacity of handling such alignments would no doubt have a significant negative effect on the alignment performance over a large-scale corpus.

### Evaluation of Alignment Result

Conventionally, precision and recall in terms of the proportion of correctly aligned sentences (or clauses) are used to measure the alignment performance. The precision is the proportion of correctly aligned pairs in all aligned pairs, and the recall is the proportion of correctly aligned pairs in

all correct pairs. For example, given a bilingual corpus of  $M$  pairs and an alignment output of  $N$  pairs among which  $A$  pairs are correct, the precision  $P = \frac{A}{N}$  and the recall  $R = \frac{A}{M}$ .  $A$  is, in a sense, the overlap of  $M$  and  $N$ .

However, since sentences and clauses are of various lengths, such percentage measures give only a rough measure about the performance. We cannot say that the correct alignment of a sentence pair of 20 words is of the same significance as that of a pair of 10 words. Thus, there seems to be a necessity to develop a more comprehensive measurement for the performance of text alignment with some finer-grained measures. For example, the proportion of words and characters in correctly aligned sentences may also give an indication of the performance in addition to the precision and recall. There is no doubt that the combination of these measures gives a more comprehensive and reliable evaluation of the performance.

In the evaluation, it is also meaningful to examine the performance against the sentence lengths and resources used. Some approaches may have a better performance on long sentences and some on short sentences. Some resources play a more important role than others.

Straightforwardly, a more comprehensive evaluation is needed for an alignment approach that carries out alignment at various linguistic levels simultaneously. For example, if an algorithm is to derive text pairs at the levels of clause, phrase and word, it need not mention that we have to look into the precision and recall at all these levels for the purpose of evaluation.

### **Example Base**

Once the text alignment phase is carried out from top to bottom at various structure levels, we actually have had an entire collection of examples in the aligned bilingual corpus. Conceptually speaking, all examples can be extracted from the aligned corpus and technically we need an example base (EB) for convenient storage and retrieval of examples. Whether it is necessary to extract all examples from the corpus is also a non-trivial issue, because an aligned bilingual corpus is itself storage for the examples. More importantly, the EB is not merely a place to pile up the examples. Rather, it needs to play the role of a language model, with the examples as structural parameters each associated with a probability estimation, for the purpose of facilitating later stage processes of EBMT, e.g., the decomposition of a source sentence into existing examples and the composition of a target sentence from a sequence of its component examples.

### Example Extraction

Extracting examples from a well-aligned bilingual text is not a trivial task, just like that given a set of scores chosen from a similarity matrix, determining the alignment is still far from trivial. Suppose that a sentence pair is word-aligned, what pairs of word sequence should be extracted as examples? For example, given the following sequences of aligned words, what examples are going to be extracted?

...	a	b	c	d	...
...	A	B	C	D	...

Word pairs, of course. And then, two-word pairs, like  $\langle ab : AB \rangle$ ; three-word pairs  $\langle abc : ABC \rangle$ ; and so on?

A serious problem with this approach is the control of example number. The number of such examples, among which many are useless, would be too large on a corpus of millions of utterances. Given  $n$  word pairs, the total number of examples as the above is in the order of  $\mathcal{O}(n^2)$ . For a corpus of  $m$  utterances with the average length  $n$ , the number of examples is in the order  $\mathcal{O}(mn^2)$ . Our computers seem to have adequate memory space to handle this complexity before  $m$  goes too large. However, a huge  $m$  is commonly encountered in current empirical approaches to human language technology, e.g.,  $m = 1,000,000$ . Assume the average utterance length  $n = 10 \sim 15$  and each example needs only 10 bytes, the space needed for the examples would be  $10mn^2 = 1.0 \sim 2.25\text{G}$  bytes. What if we have a corpus ten times larger? Notice that a larger corpus is always welcome, because it contains more knowledge that can facilitate MT. Thus, instead of relying on the computer's memory capacity, we pay more attention to what examples - if extracted - should be filtered out, in order to enhance the efficiency. Structural analysis may help us to extract only sentential constituents, to prevent too many meaningless word sequences from entering the EB. But obviously, a rigorous measure on the usefulness of an example seems necessary.

In addition to examples with continuous word sequences as the above, are there any discontinuous sequences that should also be extracted? For example, "take [something] into account" may map to something like "consider [something]" in another language. We can see that many template-like examples are quite common and useful. This appears to be another direction of research concerning example extraction: acquisition of template or

patterns from examples, where some machine learning techniques may be involved.

### **EB Management**

The EB is a crucial component in an EBMT system. It handles the storage, edition (including addition, deletion and modification) and retrieval of examples, to support the translation process, be it fully automatic or human-aided. Thus, an efficient EB must be capable of handling a massive volume of examples at an adequately high speed.

The format for literal example in the EB can be simple: a sequence of words in both the source and the target language is appropriate, although some more sophisticated language mark-up technology such as XML or RDF can undoubtedly be helpful. Also, an efficient strategy for searching through an EB of tens or even thousands of millions of example entries is necessary. In this direction, it is certainly beneficial to incorporate the machine readable dictionary (MRD) technology (Evans & Kilgarriff 1995) into the EB management.

If example patterns (or templates) and some rule-based examples bearing context-sensitive information about the preferable translation under a particular context are also considered, heterogeneous formats instead of a uniform format need to be employed in the EB for example representation and encoding. Accordingly, it is necessary to develop a more comprehensive and sophisticated EB management technology.

### **EB as a Language Model**

The EB may also play the role of a language model that needs to be utilized in later phases of EBMT to choose a suitable sequence of available examples for the translation of an input sentence and determine the word and/or chunk order in the target sentence for better readability. The former step is known as source sentence decomposition and the latter target sentence synthesis.

In a language model, structural items (e.g., the examples) are usually associated with some statistical parameters, e.g., frequencies or probabilities. In the case of n-gram model, the parameters are attached to item sequences. Frequency information is expected to incorporate not only the counts in a given bilingual corpus but also usage frequency of the examples

by users. For this purpose, the frequency information needs to be updated dynamically during translation practice.

### Example Application

The phase of example application is about how to make use of existing examples to do translation. The essence of EBMT is this: whenever you see a piece of input text that was translated before, simply use the existing translation. EBMT is about how to make use of existing translation to translate new texts following this principle. It is trivially simple to translate an input sentence that was already translated before: copy its translation to the output.

However, since the chance to re-translate a seen sentence is so dim in practice. And we also observe that what translators spend most of their effort to deal with are new sentences with translated fragments. Hence, a more critical and interesting issue in EBMT is how to translate a new sentence with fragments that are already translated.

Thus, the first task in the stage of example application is to decompose, or segment, an input sentence into a sequence of seen fragments, namely, examples. Usually, there can be more than one way to decompose a sentence in terms of a given set of known examples. Among these possibilities it is reasonable to choose the best for the next phase of translation. However, by what criterion can we say one choice is better than the other?

There can be different criteria about the goodness of a decomposition. One of the choices is probability: we segment an input sentence  $s$  into a sequence  $d(s)$  of examples  $e_1 e_2 \cdots e_n$  that is most probable in terms of a given language model. Formally put, this idea can be formulated as

$$d(s) = \arg \max_{e_1 e_2 \cdots e_n = s} p(e_1 e_2 \cdots e_n) \quad (2)$$

where the probability  $p(\cdot)$  can be computed in terms of some language model. For example, if we use a multi-gram model (of words), we have

$$p(e_1 e_2 \cdots e_n) = \prod_{i=1}^n p(e_i) \quad (3)$$

The probability  $p(e_i)$  can be approximated, in the simplest way, by its relative frequency in the example base, namely,

$$p(e_i) = \frac{f(e_i)}{N} \quad (4)$$

where  $N$  is the frequency sum of all examples in the given corpus. This is the simplest approximation. We can resort to language modeling technology for a more accurate estimation for this probability.

When the sentence decomposition is done, the next task in example application is to convert the resulting fragments from the source language into the target language. It looks like a trivial task, if there is no more than one choice of translation for a fragment in question. However, in the case of multiple translations available for a fragment, the problem appears highly similar to the one known as *word sense disambiguation* (WSD) (Yarowsky, 2000), in which the full-fledged POS tagging technology nowadays may find a critical role to play (Wilks & Stenvenson, 1998). We may hope that this technology can alleviate the problem to a great extent.

Another appropriate solution would be that we postpone the problem to the next phase, namely, target sentence synthesis, where we can pick a sequence of target fragments that has the highest *readability*, or *smoothness* of reading, among all possible sequences. This smoothness is, as we would suggest, to be measured by some probability over the sentence in question with regard to a language model, say, a multi-gram model, for the target language.

### **Sentence Synthesis and Smoothing**

After sentence decomposition and example transfer, we have a sequence of translated fragments. The next task is to combine these translated chunks into a well-formed highly readable sentence in the target language. This is recognized as the most difficult step in EBMT but “has received considerably less attention” (Sommers 1998).

Since different languages have different syntax to govern the sentential structures and word order, it won't work in most cases if we simply chain up the translated fragments in the same order as in the source language. The most critical point in this stage is to adjust the fragment order to form a smoothly readable sentence in the target language. In this sense, the sentence synthesis for EBMT is a job of smoothing for the enhancement of readability.

In general, language generation needs practical strategies for the sentence composition that consider not only the internal structure and external context of the input sentence (Collins & Cunningham 1997) but also stylistic and discourse issues with respect to culture fidelity and available text

generation techniques for MT. A set of grammar rules is surely not adequate, no matter how comprehensive the grammar would be. Remember that EBMT takes an empirical case-based knowledge engineering approach to MT. Thus, it is conceivable that a more preferable strategy for sentence synthesis is a probabilistic approach with a language model for the target language, which computes the probability for any ordering of a given set of chunks of words. Among all possible orderings, we choose the most probable one.

Given a set of translated chunks  $\{c_1, c_2, \dots, c_n\}$ , we look for the best ordering of them, as formulated below:

$$s(c_1, c_2, \dots, c_n) = \arg \max_{c'_1 c'_2 \dots c'_n \in \mathcal{O}(c_1, c_2, \dots, c_n)} p(c'_1 c'_2 \dots c'_n) \quad (5)$$

where  $\mathcal{O}(\cdot)$  denotes the set of all possible orderings of a given set of chunks. Notice that this time a linear multi-gram model does NOT work, because the chunk order has no effect on the probability of the chunk sequence, as shown in (3).

What we actually need here is a language model that is sensitive to word order in an ordering of the given chunks. The simplest choice is a fixed-order  $n$ -gram model, e.g., a bi-gram or tri-gram model. Some more sophisticated models are of course applicable, e.g., a probabilistic context-free grammar, but some more complicated NLP processing tasks such as parsing may be involved.

For example, if a tri-gram model is chosen, the probability  $p(\cdot)$  in (5) can be formulated as below accordingly:

$$p(c'_1 c'_2 \dots c'_n) = \prod_{w_{i-2} w_{i-1} w_i \in c'_1 c'_2 \dots c'_n} p(w_i | w_{i-2} w_{i-1}) \quad (6)$$

where we borrow  $\in$  to denote the “sub-string of” relation.

Another critical issue in sentence smoothing is that simply chaining up chunks in term of preferable ordering may not be enough to achieve an acceptable readability. Instead, some additional words (or chunks), such as function words, may be required for better readability. For example, when an English noun phrase is translated into Chinese, a classifier needs to be inserted. Thus, it is necessary to incorporate word insertion into the sentence generation model as given in (5), through taking into consideration a close set of smoothing words as possible chunks for generation - if adding some such chunk(s) can lead to high probability, we add them, and get a more readable sentence.

### **Other Issues**

In addition to the four stages as discussed above, there are also many other important issues involved in EBMT, for example, user interface, example filtering and pattern inference, learning and usage modeling, etc., that we have not discussed here.

As the user interface is concerned, Kay's "translation amanuensis" position still holds: translation is a kind of editing work that needs a powerful interface to give all kinds of support to an editor, including retrieving all relevant translated fragments and memorizing new translation. Memorizing new translation is also a suitable human-machine interactive approach to example acquisition. Imagine that if we have a distributed MT system for thousands of professional translators to use and input examples during their translation, what a powerful MT system we can finally produce?

In addition to the input from its users, the machine's own learning ability beyond memorization is also important. Such learning tasks include learning collocation and fixed expressions from a given corpus. There are empirical collocation extraction techniques (e.g., Smadja 1993, Smadja & McKeown 1994) that we can employ.

The other learning tasks are to infer patterns or templates from existing examples and to do example filtering for example control. There is no doubt that such learning involves difficult problems, because we are obviously in a dilemma: we need to control the example number but within a case-based learning approach no individual example should be dropped in principle. Thus, the question to ask is: how do we know which examples are useful and which ones are not? Usage frequency information can be rather useful, but a measure for the usefulness of examples is to be defined in a more theoretically-defendable way. Then, we can, hopefully, find a better strategy for example filtering.

### **Conclusions**

In general, EBMT tackles the following problem: given the translation for the fragments of a source utterance, including its words, phrases and other non-constituent chunks, infer the best choice of its translation in the target language with respect to the available translated fragments. In addition to a large-scale corpus in the target language, the translated fragments are the only resources for inference. There must be at least a sequence of fragments

that covers the entire input utterance; otherwise, the input cannot be translated completely. The fragments may or may not be adequate in number for inferring a well-formed high-quality sentence in the target language, but we do want it to be as readable as possible. To enhance the capability of translation, it is necessary to collect translated fragments from existing parallel corpora, via text alignment and example acquisition discussed above.

In the above sections, we have given an overview on the EBMT approach to machine translation. We discussed the main issues involved, including its philosophy, origins, history and the four stages of work involved. We consider EBMT as an empirical case-based knowledge engineering approach to MT, in which the major means for knowledge acquisition is example acquisition by text alignment from large-scale parallel bilingual corpora. We reviewed the current state of the text alignment technology and recognized the importance of performing alignment at various linguistic levels for the acquisition of a comprehensive collection of examples. We also discussed some critical ideas about how to apply existing examples to translate new utterances, via source sentence decomposition and target sentence synthesis with the aid of language models.

The underlying principle for EBMT is as simple as this: remember every thing translated in the past and use everything available to facilitate the translation of the next utterance. We know computers are the most fantastic machines to memorize such things as text pairs and their frequencies, and we thus have reason to believe that EBMT represents the MT approach with the greatest potential.

## **Acknowledgements**

This work is part of the research output from the CERG project “EBMT for HK Legal Texts” funded by HK UGC under the grant #9040482, with Jonathan J. Webster as the principal investigator and Chunyu Kit, Caesar S. Lun, Haihua Pan, King Kuai Sin and Vincent Wong as investigators. The authors wish to thank all team members who have contributed to the research work that enables this paper. Correspondence concerning this work should be addressed to Dr. Jonathan J. Webster, CTL, CityU of HK, Tat Chee Ave., Kowloon, HK.

## References

- Arthern, P. J. (1978). Machine translation and computerized terminology systems: a translator's viewpoint. *Translating and the Computer: Proceedings of a Seminar* (pp. 77–108). London.
- BLIS (1998). Bilingual laws information system (BLIS). Info. Tech. and Resources Unit, Admin. Division, Dept. of Justice, HK Government. Information available from <http://www.justice.gov.hk/>.
- Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lefferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16, 79–85.
- Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. *ACL-91* (pp. 169–76). Berkeley.
- Brown, P. J., Pietra, S. A. D., Pietra, V. J. D., Lefferty, J. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19, 263–311.
- Brown, R. D. (1996). Example-based machine translation in the Pangloss system. *COLING-96* (pp. 169–174).
- Chen, S. (1993). Aligning sentences in bilingual corpora using lexical information. *ACL-93* (pp. 9–13). Columbia, Ohio.
- Church, K. (1993). Char-Align: A program for aligning parallel texts at the character level. *ACL-93* (pp. 1–8). Columbia, Ohio.
- Collins, B., & Cunningham, P. (1995). A methodology for example based machine translation. *CSNLP-95: 4th Conf. on the Cognitive Science of NLP Proceedings*. Dublin.
- Daga, I., Church, K. W., & Gale, W. A. (1993). Robust bilingual word alignment for machine aided translation. *Proc. of the Workshop of Very Large Corpora* (pp. 1–8). Columbus, OH.
- Evans, R., & Kilgariff, A. (1995). MRDs, standards and how to do lexical engineering. *Proc. of 2nd Language Engineering Convention* (pp. 125–32). London.
- Fung, P., & McKeown, K. (1997). A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12, 53–87.
- Gale, W. A., & Church, K. W. (1991). A program for aligning sentence in bilingual corpora. *ACL-91* (pp. 177–84). Berkeley.
- Hutchins, W. J. (1998). The origins of the translator's workstation. *Machine Translation*, 13, 287–307.

- Kay, M. (1997). The proper place of man and machines in language translation. *Machine Translation*, 12, 3–23. First print as research report CSL-80-11, Xerox PARC, Palo Alto, CA., 1980.
- Kay, M., & Roscheisen, M. (1993). Text translation alignment. *Computational Linguistics*, 19, 75–102. First print as Technical Report P90-00143 in Xerox Palo Alto Research Center in 1988.
- Kit, C., & Webster, J. J. (1992). Machine translation of idioms based on tokenization. *Proceedings of 1st Singapore International Conference on Intelligent Systems*. Singapore.
- Melby, A. (1995). *The possibility of language: A discussion of the nature of language*. Amsterdam: John Benjamins.
- Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji (Eds.), *Artificial and human intelligence*, 173–180. Amsterdam: North-Holland.
- Nirenburg, S., Beale, S., & Domashnev, C. (1994). A full-text experiment in example-based machine translation. *Int'l Conf. on New Methods in Language Processing* (pp. 78–87). Manchester.
- Nirenburg, S., Domashnev, C., & Grannes, D. (1993). Two approaches to matching in example-based translation. *TMI'93* (pp. 47–57). Kyoto.
- Pan, H. H. (1986). A machine translation system for scientific titles of English. *Proc. of 1986 Int'l Conf. on Chinese Computing*. Singapore.
- Pan, H. H. (1987). Towards understanding-based MT. *Proc. of the Int'l Conf. on Chinese Information Processing*. Beijing.
- Pappegaaïj, B. C., Sadler, V., & Witkam, A. P. M. (1986). *Word export semantics: An interlingual knowledge-based approach*. Dordrecht: Reidel.
- Sato, S. (1993). Example-based translation of technical terms. *TMI'93* (pp. 58–63). Kyoto.
- Sato, S., & Nagao, M. (1990). Toward memory based translation. *COLING-90* (pp. 247–252). Helsinki.
- Schubert, K. (1986). Linguistic and extra-linguistic knowledge: a catalogue of language-related rules and their computational application in machine translation. *Computer and Translation*, 1, 125–152.
- Sin, K. K., & Roebuck, D. (1996). Language engineering for legal transplantation: conceptual problems in creating common law Chinese. *Language and Communication*, 16, 235–254.
- Smadja, F. A. (1993). Retrieving collocation from text: Xtract. *Computational Linguistics*, 19, 143–177.

- Smadja, F. A., & McKeown, K. R. (1994). Translating collocation for use in bilingual lexicon. *Proc. of the ARPA Human Language Technology Workshop*. Princeton, N.J.
- Somers, H. L. (1998). “new paradigms” in MT: the state of play now that the dust has settled. *Machine Translation Workshop, ESSLLI-98* (pp. 22–33). Saarbruecken, Germany.
- Somers, H. L. (2000a). Example-based machine translation. In R. Dale, H. Moisl and H. Somers (Eds.), *Handbook of natural language processing*, 611–627. New York: Marcel Dekker.
- Somers, H. L. (2000b). Machine translation. In R. Dale, H. Moisl and H. Somers (Eds.), *Handbook of natural language processing*, 329–346. New York: Marcel Dekker.
- Sumita, E., Furuse, O., & Iida, H. (1993). An example-based disambiguation of prepositional phrase attachment. *TMI'93* (pp. 80–90). Kyoto.
- Sumita, E., Iida, H., & Kohyama, H. (1990). Translating with examples: A new approach to machine translation. *TMI'90* (pp. 203–212). Texas.
- Webster, J. J., & Kit, C. (1992). Tokenization for machine translation: What can be learned from Chinese word identification. *Proc. of 3rd Int'l Conf. on Chinese Information Processing*,. Beijing.
- Wilks, Y., & Stenvenson, M. (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic discrimination. *Natural Language Engineering*, 4, 135–144.
- Wu, D. (1995a). Grammarless extraction of phrasal translation examples from parallel texts. *TMI'95* (pp. 354–372). Leuven, Belgium.
- Wu, D. (1995b). Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. *IJCAI-95* (pp. 1328–1335). Montreal.
- Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23, 377–404.
- Wu, D. (2000). Alignment. In R. Dale, H. Moisl and H. Somers (Eds.), *Handbook of natural language processing*, 415–458. New York: Marcel Dekker.
- Wu, D., & Xia, X. (1995). Large-scale automatic extraction of an English-Chinese translation lexicon. *Machine Translation*, 9, 285–313.
- Yarowsky, D. (2000). Word-sense disambiguation. In R. Dale, H. Moisl and H. Somers (Eds.), *Handbook of natural language processing*, 629–654. New York: Marcel Dekker.